ISSN:0975-9646

User-Centric Based Crawler System

Santosh Kumar Uppada

CSE Department,
MVGR College Of Engineering, JNTU Kakinada
Chintalavalasa,Vizianagaram.

***Abstract*— The motto of researchers is to pose technical endeavours in a propounded manner. There raises a strong impact of the pre-existing slogs made by other researches. Certain web crawlers have been formulated to extract quondam aspects that exist over net. Demand rises for designing certain classifiers which interacts with multiple hosts. A robust, flexible and manageable crawler system is to be coined which crawls data on par of user interest. This paper poses a methodology which could crawl through data repositories required as per the researcher interest. Certain classification algorithm like C4.5 has been proposed to classify the patterns which sound to be interesting while mining data from huge data repositories available on net.**

***Keywords*— classification, crawler, C4.5, Reference Comment Crawler (RCC), Quinlan ID3, decision tree, statistical classifier, predictor, gain ratio.**

I. INTRODUCTION

Researchers generally present their technological endeavours by means of technical papers. These writings are generally meant for global level acceptance of any advancement that one claims to present. There are certain conventions that have been developed for presenting the research work by first presenting "Related Work", which comprises of the aspect of comparing and contrasting their work with the work that has already being brought to pass. Citation is generally being used to index the resources from which the present work has been derived. Certain pre-statements and post-statements are used to originate the idea from different related papers. Commented sentences are used, which are equipped with a reference identifier which defines the positional values of the research references. Different journal and conference papers of different periods are being studied from different hosts, which serve as repositories of the research works.

The advent improvement in the field of internet has made all the information to come into one's hand. Information is being shared and presented in different forms and at different levels. In general, search engines are being used in searching for particular papers being cited in the bibliography section. Immense web source searching is being used for extracting repository elements. This process generalises and reduces the time and effort that takes for searching the previous works in the propounded field with more feasibility metric.

In addition to the central server repositories, there are many intermediate bots that are being defined for systematic browsing from net. These bots are generally used for web indexing. These bots generally represent web-browsers which aim at collecting information that is most likely being requested by major sector of people. A general Web-Crawler may also be termed as Web Spinder, an ant, an automatic indexer, or as a Web Scutter. In general, web search engines and other related sites utilize the endeavour of this spindering or crawling software for updating and indexing the content with respect to other sites. Crawlers are also well-known for validating hyperlinks and HTML Code.

In general researchers aim at providing reference comments on the papers that has been already cited. Comments are very much essential as researcher aims at proposing the pros and cons of the study, in accordance to the feasibility. Diversifying comments may exist for the same paper as the viewing ability generally differs from different perspective.

Reference Comments Crawler (RCC) systems are generally developed that aims at gathering reference points and comments pertaining to a particular study. In general Reference paper crawling and comment sentence extraction are the major functions related to the RCC systems. Crawler, filter and analyzer comprises the major part of the system.

When there exist multiple reference papers existing for a particular author, year of publication is also taken as a promising metric. Most of the cases used only publisher name with worst cases in which the month and year of publication details are also taken as a key metric for extracting information or comments cited by various researchers working for the same domain.

The process of extraction of data with prior knowledge on the class type helps in gathering more accurate data in more promising manner. Classification is claimed to be the methodology where group members are used for predicting group members for given data instances. Data that has been computed is segregated into training and test samples which are used for classifying and predicting the outcomes.

There have been many classification algorithms that have been used at a very promising level in order to group members that have been collected. Sets has been maintained under the name of training and test sets, in which first an algorithm has been proposed by utilizing the same for evaluating the same with the other set of data termed to be used for testing. Decision trees are the most advantageous methods that are been developed with respect to the efficiency in terms of understandability, scalability, classification extent, and high speed. Quinlan's ID3 has

been made as a benchmark for improvisation in the construction of decision trees in the next generation C4.5, C5.0 algorithms.

The present research has been constructed over classifying different data objects from the research papers. Process deals with analysing research work with respect to any of author names, area of interest as cited in profile, research domains, publication bodies, indexes regarding citation levels. The variant to the primitive C4.5 has been used in order to classify the data that has been extracted and crawlers are been utilized on the same which extract the imperative information which helps in presenting new methodologies and comparing the same with the previously defined, in a well propounded manner.

II. LITERATURE REVIEW

A great research has been proposed in terms of extracting information over internet. The aspect of writing a research paper is no way an easy task and claims more background knowledge. Issue rises with respect to attaining information regarding the previously dealt methodologies, which serves as a building block for proposing new methodology [1].

Search engines originate in finding useful information by extracting information from different web resources. Web Spinder, termed as a vital elementary of search engine, claims finding and collecting useful information from different web repositories. Content Graph has been claimed as a primitive method in building a crawler basing on specific topic. Real web link relations are being used in developing this crawler, which captures link hierarchies and extracts information basing on distance measured from off-topic to the target pages. Seed pages are being extracted at first, using which an associated context graph is defined. Focused crawlers are used for classifying downloaded pages to form a queue of different classes. Crawler is trained for retrieving information from unvisited pages from high priority queues, meant for crawling [2]. Breadth-first crawlers focussed with cosine similarity, termed as CCG, a hierarchical graph of concepts, and an attribute based FCA mechanism used in retrieval precision after updating CCG [3].

Development of strategies in crawling only relevant data, designing parallel crawlers and restricting of hypertext documents are termed as other areas of development. Parallel crawler based on augmented hypertext documents (PARCAHYD) is being derived for making crawlers, work in parallel. The first part of system claims in dividing document retrieval system into crawling system and hypertext (augmented) document system, wherein Table of Links(TOL) are derived for the extraction of linked documents, housed on external sites. Documentation, mapper and crawl worker level are defined in aspect of parallelization. These remove the bottlenecks at document level. Schemes for managing volatile information and reduction of redundant data at repository level are employed by all crawlers [4].

Major challenge comes with aspect downloading multiple pages, maintaining its quality and in reducing bandwidth consumption. Incremental crawlers are used in crawling web pages continuously. Focused crawlers are maintained in dealing with maintenance of freshness of databases majorly when crawlers are maintained with missing relevant pages. Distributed crawlers are maintained for partitioning the collection, distribution of URL agents, partitioning and balancing the load of web servers, reducing bandwidth and effective cache design. Mobile crawlers are other variant to the basic crawlers, aimed to work on security, non-availability of required information and in improvisation of mobile crawling algorithms [5]. Scalability is the major issue that has been stated using distributed crawler system. Simulation testbed consisting of several workstations, simulating web using either of pages generated artificially or stored partially as a snapshot of web. High performance network systems are proposed as next generation testbed crawlers [6].

With respect to web based crawler algorithms, there have been many methodologies that are being proposed which define the efficiency of the bots that are meant for automatically extract information from hierarchy of relevant papers. Freshness and re-visiting are the major components for dealing. Methods like Breadth first search, Depth first search, Page Rank, Genetic, Naive Bayes classification and HITS algorithms are meant for automatically extract information. Some of these methods even work for artificially generated, large distributed and interconnected pages and in calculation of various measures which promotes for extracting strong and precise information [7].

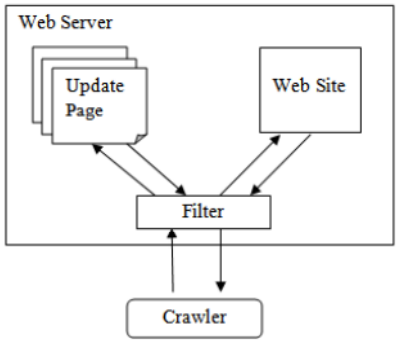

Figure1. Filtering of pages from websites basing on LAST-VISIT

The advent of Query based algorithms has made the process of information, quite easier. As depicted in the figure1, web server pages are being filtered depending upon certain measuring factors. LAST_VISIT is the major part taken into account, which deals with last crawling time of particular web-site. Filters used helps in updating the web-servers periodically depending on the visit ratio. Crawlers are aimed to the visit the updated pages as updated in the servers with respect to the URLs. This process also works as unethical crawling method [8].

This aspect of filtering and dynamic addition of URLs and lack of efficient refreshing techniques are making the present crawlers to impulse false alarms and unnecessary traffic. Frequency in this aspect helps in calculating refresh time dynamically. The computation of refresh time helps in optimizing visits by dynamically

assigning to sites. Efficiency of crawling system is improved by revisiting frequency for a site. Interest of users serves as an vital part for crawling pages with more speed than those pages which are rarely surfed by the users [9].

The major task of the localized crawlers deals with the aspect of retrieving information behalf of search engine. Mobile crawlers reduce HTTP overhead by transferring the crawler to the source of the data. Crawler Managers, Remote page selection and Remote page Filtering helps in accessing content that is relevant. Crawler managers, Query engines, Database drivers related to connection and command managers, Bots exclusion protocol and parallel implementations dealing with DNS resolvers, URL handlers and crawler migrations are the major components of migrating web crawlers, in parallel [10].

The major issue rises with the lexical data as it consists of synonyms, Holonyms, Meronyms, and Antonyms of particular language. Synset are generally maintained for particular word. Crawlers are to be enhanced in dealing such aspects. URL crawler and searching modules are derived in this aspect. Poor indexing terms are generally omitted using stop words. Stemming is another promising aspect dealt which retrieves different forms for a given word. Precession and recall has been significantly balanced for crawling relevant data. Domain ontology, Description and Information Retrieval algorithms add efficiency for the standing crawlers [11].

Seeding of URLs is another variant to the crawler systems for deriving more relevant data. Topic keyword is used for different search engines for information extraction. Relevancy score is calculated by using link weight, text content similarity using Levenshtein distance measure or probability method. This method of dealing with topic keywords- using similar and dissimilar methods makes the method of extracting pages more lucid [12]. Crawling process starts with crawling from the list of seed URLs related to specific topic. URL distillers are managed for relating seeded URL for particular topic-search URL. C-proc processes are used for such phenomenon [13].

In addition to these URL seeding, crawlers depending on the comments are also been proposed. Reference Comments Crawler System (RCC) has been proposed for collecting comments about the reference papers. RCC in general consists of Reference Paper Crawling (RPC) and Comment Sentence Extraction (CSE). The RPC searches for papers that have citations about the reference papers listed in the reference section. Title of the reference paper is taken as an input. The RCC parser analyzes and finds the papers having citations for reference papers. Comment sentence extraction starts with the process of Reference Identifier Extraction, which deals with extraction of reference citations given in any format. Precision and Recall are preserved. Comment sentences can be taken under category of PreSentences and PostSentences. Precision in certain cases relates to relatively lower values because of the reason that the reference identifier stated are different from the description in the body of the paper. If same person reference is cited for times, year and sometimes with its combination is used in citation of particular papers. Recall measure also counts to be low when there is either mismatch in representation of citations or identifiers are given as subscripts in the body of the paper [14].

Classification algorithms have attracted attention in fields of machine learning and other data mining aspects. These are used for assigning items of collection to a target category or class. Classification beings with analysing class members whose class assignments are known. Classification models are developed by comparing the predicted values to known target values for finding relations. Decision trees are constructed as a predictive model which maps observed values to the targeted value. In these tree structures, leaves represent class labels and branches represents conjunctions of features relating to different labels. Decision trees are the most advantageous methods that are been developed with respect to the efficiency in terms of understandability, scalability, classification extent, and high speed. Quinlan's ID3 has been made as a benchmark for improvisation in the construction of decision trees in the next generations.

A major invariant to classification has been using C4.5, which is basically a statistical classifier. This tree at each node chooses attribute that most likely to split the set of samples into subsets relating to one or more classes. The split criterion is calculated depending on the information gain, which is the difference measure in the entropy. Attribute which has higher information gain value is being chosen for decision making. This algorithm is rebuilt on the thus formed subset [15].

Crawler depending on URL seeding and comment based approaches have been proposed on the basis of this classification algorithms. Crawlers with respect to decision trees or classification are generally meant for predicting whether a page is related to the aspect of search. Prediction can be done basing on domain, author or publication body. Link focused crawlers focus on the surrounding text around the links that the seeding URL is pointing. Classification rules are maintained depending on the building factor of the decision tree [16].

Query based classification is maintained for searchable web databases. Classification helps in extracting information about the uncrawlabe content. This method reaches to correct classification decision quickly, while crawling based technique depends on topic distribution. Classification accuracy and efficiency is being uplifted by this methodology [17].

III. PROPOSED METHODOLOGY

In general data analysis is being worked on basis of either classification or prediction. These are used for deriving important classes for describing the class functions or in predicting future outcomes. Analysis is categorized either on basis of classification or prediction. These are taken under the prospect of supervised learning where class names are pre-defined and the outcome is being checked against certain amount of data that has been extracted which is used for defining a particular algorithm. Here it should be remembered that the other part of data is being used for checking the correctness of algorithm that has been derived.

Classification corresponds to different aspects defined on data and maintenance of class names for the pages derived. Decision trees are generally being produced on par of data that has been extracted.

C4.5 is generally used on par of developing a classification tree. It is depicted by Ross Quinlan. It has been coined as advancement to the pre-existing decision tree algorithm, ID3. In general, C4.5 is regarded as a statistical classifier.

C4.5 has been defined as an extension to the pre-existing ID3 algorithm. While working with C4.5, in building a decision tree, one can deal with training sets that have records with unknown attributes. Process is done by evaluating gain ratio, for attribute by considering the records where the attributes are defined.

While constructing decision tree, records can be classified with unknown attribute values by estimating the probability of various possible results.

C4.5 generally yields to three different base cases.

Case 1. It confines to the case where all the samples of the entire list, belongs to the same base class. C4.5, in this case, simply creates a leaf node of the decision tree, which makes the class to be chosen readily.

Case 2. It confines to the case where Information gain is not formed by any of the features that has been tracked. In this case, C4.5 tends to create a decision node higher up to the tree structure depending on some expected value from the class predictive.

Case 3. Ii confines to the case where an instance of previously unseen class has been encountered. In this case, C4.5 tends to create a decision node higher up to the tree structure depending on some expected value from the class predictive.

Input: An attribute-valued dataset T

Output: Decision tree based on attribute set.

Assumption:

a. Threshold value is being derived by sorting values if attributes are numerical in type.
b. Instantaneous value is being taken directly from class if the attributes considered are of nominal type.

Steps:

1. Initially, take Tree = { }
2. If T is pure, or if there is some limit has been met, then
3. Terminate
4. endif.
5. for all attribute t ε T do
6. Compute information-theoretic criteria if we split on a
7. end for
8. t_{high} = Best attribute according to above computed criteria
9. Tree = Create a decision node that tests t_{high} in the root
10. T_v = Included sub-databases from T based on $t_{high.}$
11. for all values of T_v do
12. $Tree_v$= C4.5(T_v)
13. Attach $Tree_v$ to the corresponding branch of Tree
14. end for
15. return Tree.

IV. RESULTS AND CONCLUSION

The process of developing crawler system based on classification starts with extracting certain information about the research work. Certain systems can be used which are solely dedicated for extracting research work that has already been laid. Certain references are being extracted and depending on the criteria like h-index of h-median values, the total pages are being segregated.

Once the set has been derived, in order to process the data with ease, the data related to publications, domains are being defined with certain codes relating to its importance. Every attribute has been given with short notations. Sample of the above data has been taken in order to classify the content derived. Tanagra tool is being used for processing.

Table1. Instance of the synthetic data for processing

domain	publication	h5-index	h5-median	true_class	pub_pwr	h5-bst
1	p1	69	113	yes	high	yes
2	p2	21	83	no	low	no
3	p3	54	82	yes	high	yes
4	p4	36	62	yes	high	yes
5	p5	36	47	yes	high	yes
6	p6	35	49	no	high	no
7	p7	32	48	no	low	yes
8	p8	27	49	no	low	no
9	p9	27	40	no	low	no
10	p1	37	34	yes	high	yes
11	p2	22	54	no	low	no
12	p1	35	12	yes	low	yes
1	p4	32	34	yes	high	yes
2	p1	30	23	yes	high	yes
5	p3	55	5	yes	high	yes
1	p2	21	4	no	low	no
1	p5	45	33	yes	high	yes

Steps illustrating the processing in Tanagra

1. The Tanagra tool has been opened, the data image is being extracted. It should be remembered that either of csv, plain text or arff files have been given as an input, once defied. Discrete and continuous attribute vales are also be taken.
2. Feature selection is being carried out by selecting "Define Status" palette. Attributes that are being used for classifications are also being used.
3. Now from SPV Learning, C4.5 algorithm has been selected. Before working on the respective classifier, status of the classifying attribute has been defined from "Feature Selection" palette.

Scheme: TANAGRA.classifiers.trees.C4.5 -C 0.25 -M 2
Relation: santu1
Instances: 24
Attributes: 7
domain
publication
h5-index
h5-median
true_class
pub_pwr
h5-bst
Test mode:10-fold cross-validation

Figure1 . Figure illustrating attributes and fold count for classification

The above figure illustrates the no of folds that have been provided. Domains once selected Are being send for processing using certain training set.

The above figure illustrates the no of folds that have been provided. Domains once selected Are being send for processing using certain training set.

```
C4.5 pruned tree
true_class = yes
|  h5-index <= 45
|  |  h5-median <= 23: p1 (2.0)
|  |  h5-median > 23
|  |  |  h5-index <= 36: p4 (3.0/1.0)
|  |  |  h5-index > 36: p1 (2.0/1.0)
|  h5-index > 45: p3 (3.0/1.0)
true_class = no
|  h5-index <= 22: p2 (3.0)
|  h5-index > 22
|  |  domain <= 7: p6 (2.0/1.0)
|  |  domain > 7: p8 (2.0/1.0)
Number of Leaves  :  7
Size of the tree :  13
Time taken to build model: 0 seconds
```

Figure2. Training set classifier model

Training sets are being designed. Once the training set classification has been illustrated, cross-validation issues are being handled as depicted in figure.

Correctly Classified Instances 11.7647 %	2
Incorrectly Classified Instances 88.2353 %	15
Kappa statistic	-0.02
Mean absolute error	0.1869
Root mean squared error	0.3709
Relative absolute error 93.5274 %	
Root relative squared error 115.7009 %	
Total Number of Instances	17
Ignored Class Unknown Instances	7

Figure3. Stratified cross-validation

Table 2. Detailed Accuracy By Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0	0.308	0	0	0	0.281	P1
0.667	0	1	0.667	0.8	0.778	P2
0	0.133	0	0	0	0.5	P3
0	0.2	0	0	0	0.705	P4
0	0.067	0	0	0	0.205	P5
0	0.188	0	0	0	0.304	P6
0	0.063	0	0	0	0.304	P7
0	0.063	0	0	0	0.283	P8
0	0	0	0	0	0.283	P9
Weighted Avg.	0.118	0.138	0.176	0.118	0.141	0.438

Once a specified classification has been illustrated, major potential elements like precision, Recall, FP-Rate, F-Measure, class values are being defined. The resultant which is taken as confusion matrix is being derived.

```
a b c d e f g h i   <-- classified as
 0 0 1 2 1 0 0 0 0 | a = p1
 0 2 0 0 0 0 0 1 0 | b = p2
 2 0 0 0 0 0 0 0 0 | c = p3
 2 0 0 0 0 0 0 0 0 | d = p4
 0 0 1 1 0 0 0 0 0 | e = p5
 0 0 0 0 0 0 1 0 0 | f = p6
 0 0 0 0 0 1 0 0 0 | g = p7
 0 0 0 0 0 1 0 0 0 | h = p8
 0 0 0 0 0 1 0 0 0 | i = p9
```

Figure4. Confusion Matrix

Once the confusion matrix has been derived, properties are being derived by classifying the entire data set that has been extracted. Decision tree is being constructed for classifying the data set that has already being defined.

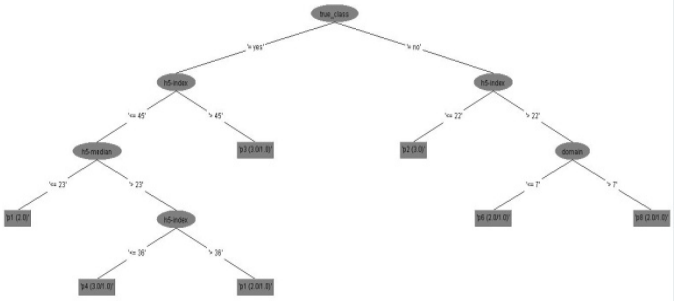

Figure5. Decision Tree constructed using C4.5 Classifier

Figure shows the decision tree which has been formulated when publication is being taken as a class variable. The leaf tree is being used to predict the outcomes from the synthetic data.

V. CONCLUSION

Thus the proposed system has succeeded in classifying the content that has been derived. Decision trees are being constructed basing on the limit parts that have been taken for each of the data that is being extracted. Even missing values are being handled using the proposed system.

There has a major demand of posing different classifiers that could relate the previous work to formulate certain new endeavours. Decision tree as illustrated is a formulating block that could segregate data properties as dependent on the user interest. Paper has made the concept of developing a crawler which could be used for retrieving information in a more propounded manner. Publication is being taken as a class label to classify the data that has been considered. This could be used for prediction analysis.

REFERENCES

[1] S.S. Dhenakaran and K. Thirugnana Sambanthan, WEB CRAWLER - AN OVERVIEW, International Journal of Computer Science and Communication Vol. 2, No. 1, January-June 2011, pp. 265-267. 2011

[2] Diligenti M, Coetzee F M, Lawrence S et al. Focussed crawling using context graphs. In: proc of the international conference on very large database (VLDB). 2000.

[3] Zhaoqiong GAO, Yajun DU, Liangzhong YI, Yuekui YANG, Qiangqiang PENG, Focused Web Crawling Based on Incremental Learning, Journal of Computational Information Systems6:1(2010) 9-16. 2010.

[4] A.K Sharma, J.P Gupta, D.P Agarwal, PARCAHYD, An Architecture of a parallel crawler based on Augmented Hypertext documents, International Journal of Advancements in Technology, ISSN 0976-4860. 2010.

[5] Satinde r Bal Gupta, The Issues and Challenges with the Web Crawlers, International Journal of Information Technology & Systems, Vol. 1; No. 1: ISSN: 2277-9825. 2011.

[6] Vladislav Shkapenyuk, Torsten Suel, Design and Implementation of a High-Performance Distributed Web Crawler, ACM Transactions on Internet technologies. 2011.

[7] Pavalam S M, S V Kashmir Raja, Felix K Akorli and Jawahar M, A Survey of Web Crawler Algorithms, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 1, November 2011. 2011.

[8] S. S. Vishwakarma, A. Jain, A K Sachan, A Novel Web Crawler Algorithm on Query based Approach with Increases Efficiency, International Journal of Computer Applications (0975 – 8887) Volume 46– No.1, May 2012. 2012.

[9] Niraj Singhal, Ashutosh Dixit, R. P. Agarwal, A. K. Sharma, Regulating Frequency of a Migrating WebCrawler based on Users Interest, International Journal of Engineering and Technology (IJET). 2012.

[10] Abhinna Agarwal, Durgesh Singh, Anubhav Kedia Akash Pandey, Vikas Goel, Design of a Parallel Migrating Web Crawler, International Journal of Advanced Research in Computer Science and Software Engineering,Volume 2, Issue 4, April 2012.

[11] Anthoniraj Amalanathan , Senthilnathan Muthukumaravel, Semantic Web Crawler Based on Lexical Database, IOSR Journal of Engineering Apr. 2012, Vol. 2(4) pp: 819-823. 2012.

[12] S. Subatra Devi1, Dr. P. Sheik Abdul Khader, Topic-specific Web Crawler using Probability Method, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 13, Issue 1 (Jul. - Aug. 2013), PP 102-106. 2013.

[13] Rohith Kumar, Virendra Kumar, Savita Shiwani, Dinesh Goyal, An Extended model of Topic Driven Focused Crawler using Parallel Crawler, International Journal of Advanced Research inComputer Science and Software Engineering, Volume 4, Issue 6, June 2014.

[14] Hocheol Jeon Agency for Defense Development, Korea, A Reference Comments Crawler for Assisting Research Paper Writing, The International Arab Journal of Information Technology, Vol. 11, No. 5, September 2014.

[15] Salvatore Ruggieri, Efficient C4.5, IEEE transactions on knowledge and data engineering, Vol. 14, No.2, 2002.

[16] Caliskan, K. ; Dept. of Comput. Eng., Turgut Ozal Univ., Ankara, Turkey ; Ozcan, R., Comparing classification methods for link context based focused crawlers, Electronics, Computer and Computation (ICECCO), 2013 International Conference. 2013.

[17] Luis Gravano, Panagiotis G. Ipeirotis, Mehran sahami, Query-vs. Crawling-based classification of searchable web databases, IEEE bullietin. 2013.