Searching and Classification of List of Keywords and URLs Using SVM Classifier

J.Santhiya, K.S Bhuvaneswari

Abstract— Internet forums are in the main used for discussions where users can request and exchange data with others. Forum contains form of topics associated with user interest. The goal of net travel is to retrieve forum content with smallest overhead. The target of the net crawlers is to send users from entry page to thread page. The thread page contains data content relevant to user question. The projected net travel primarily based on the mechanism that supports multiple keyword based retrieval using ontology. Ontology helps the crawler to extract and aggregate information from a specific domain. This proposed methodology addresses the uniform resource locator kind recognition drawback to get rid of the duplicate and unwanted pages. The strategy is evaluated exploiting SVM classifier and located to be higher than existing systems.

Index Terms—Crawlers, Ontology, SVM Classifier

.

1 INTRODUCTION

Web forums are mainly used for discussion among users and to extract useful information from it For example the Thyroid Forum is a place where people can raise query related to it and get answers. The forums contain a rich variety of information that is unstructured or structured data [11], Q&A pairs [5], ranked product features that are grouped using opining mining from forum posts. To get an idea from forum their content must be downloaded first, some crawler adopts breadth first crawling strategy but it is ineffective and inefficient because it visits from root node at each time of crawling. Many duplicate links that point to a common page with different URL. Normal crawler blindly follow these links to crawl many duplicate pages which makes the crawlers ineffective. iRobot, a web crawler uses breadth first crawling to crawl web pages and it is found to be 47% effective.

These problems can be solved by using supervised forum crawler to remove unwanted and duplicate pages. The idea is to crawl relevant content with smallest overhead. It navigation path from entry page to thread page. Links are provided between the pages to find its destination pages. Links between an entry page and an index page or between two index pages are referred as index URLs. Links between an index page and a thread page are referred as thread URLs. Links connecting multiple pages of a board and multiple pages of a thread are referred as page-flipping URLs.A crawler starting from the entry URL, it identifies the index URL, thread URL, and page-flipping URL. These URLs regular expression pattern is learned by ITF Regexes. Entry URL is varied from forums to forums. So find the entry URL is first. This supervised crawler is compared with generic breadth first crawler, iRobot crawler [3], structure driven crawler [11], the supervised crawler outperforms these crawlers in terms of effectiveness and coverage.

The multi keyword web crawling can be used for specific area to retrieve the particular content. It deals with ontology used for finding similarities between the keywords. Based on the r keyword search it should be matched with some relevant content. All of the matching content should not be retrieved. Using the rank method, content should be retrieved. Ranking is based on the highest similarities between the matching keyword content. It uses the support vector machine to predict the data .Radial Basis Function kernel SVM used to measure the similarities between the examples.

Compared linear kernel SVM with radial basis function kernel SVM, linear kernel SVM does not perform high dimensionality in large text data set. So we move in to RBF kernel to achieve the high accuracy and data should identified as it should be correct or not. The same process in the existing can be used for thread page identification by using RBF kernel to increase the data should retrieve in particular domain.

2 Literature Review

2.1 UNDERSTAND THE CONTENT AND THE STRUCTURE OF A FORUM

iRobot, which has intelligence to understand the content and the structure of a forum site, and then decide how to choose traversal paths among different kinds of pages. To do this, we first randomly sample a few pages from the target forum site, and introduce the page content layout as the characteristics to group those pre-sampled pages and re-construct the forum's sitemap. After that, select an optimal crawling path which only traverses informative pages and skips invalid and duplicate ones.

The extensive experimental results on several forums show the performance of our system in the following aspects: 1) Effectiveness – Compared to a generic crawler, iRobot significantly decreases the duplicate and invalid pages; 2) Efficiency – With a small cost of pre-sampling a few pages for learning the necessary knowledge, iRobot saves substantial network bandwidth and storage as it only fetches informative pages from a forum site; and 3) Long threads that are divided into multiple pages can be re-concatenated and archived as a whole thread, which is of great help for further indexing and data mining. iRobot learns URL information to discover new in crawling, but URL location might be invalid when page structure changes.

2.2 URL DE-DUPLICATION

They present a set of techniques to mine rules from URLs and utilize these rules for de-duplication using just URL strings without fetching the content explicitly. This technique is composed of mining the crawl logs and utilizing clusters of similar pages to extract transformation rules, which are used to normalize URLs belonging to each cluster. Preserving each mined rule for de-duplication is not efficient due to the large number of such rules. We present a machine learning technique to generalize the set of rules, which reduces the resource footprint to be usable at web-scale. The rule extraction techniques are robust against web-site specific URL conventions. It compares the precision and scalability of our approach with recent efforts in using URLs for de-duplication. Supervised crawler is on efficient and large-scale deduplication of documents on the WWW.

Web pages which have the same content but are referenced by different URLs are known to cause a host of problems. Crawler resources are wasted in fetching duplicate pages, indexing requires larger storage and relevance of results are diluted for a query. The presented techniques to scale pair-wise Rule generation and introduced a decision tree based Rule generalization algorithm. In this research we build on that work by extending the URL and Rule representations and introduce algorithm for finding host specific delimiters. Together these set of techniques form a robust method for de-duplication of web pages using URL strings. While the URL and Rule representation is complete and covers most patterns, we consider two additional patterns due to their significance on the Web: Deep Token components in a URL and URL component alignment. Propose a technique for extracting host specific delimiters and tokens from URLs. We extend the pair wise Rule generation to perform source and target URL selection. And also introduce a machine learning based generalization technique for better precision of Rules. Collectively, these techniques form a robust solution to the deduplication problem.

Another related work is near-duplicate detection. Forum crawling also needs to remove duplicates. But content based duplicate detection [6], [9] is not bandwidth efficient, because it can only be carried out when pages have been downloaded. URL-based duplicate detection [4], [7] is not helpful. It tries to mine rules of different URLs with similar text. However, such methods still need to analyze logs from sites or results of a previous crawl. In forums, index URLs, thread URLs, and page-flipping URLs have specific URL patterns. Thus, in this paper, by learning patterns of index URLs, thread URLs, and page-flipping URLs and adopting a simple URL string de-duplication technique (e.g., a string hash set), FoCUS can avoid duplicates without duplicate detection.

3 Web Crawler

Web Crawler is a meta search engine that blends the top search results from Google Search and Yahoo Search. WebCrawler also provides users the option to search for images, audio, video, news, yellow pages and white pages. A web crawler (also known as a web spider or web robot) is a program or automated script which browses the World Wide Web in a methodical, automated manner. There are several uses for the program, perhaps the most popular being search engines using it to provide webs surfers with relevant websites.

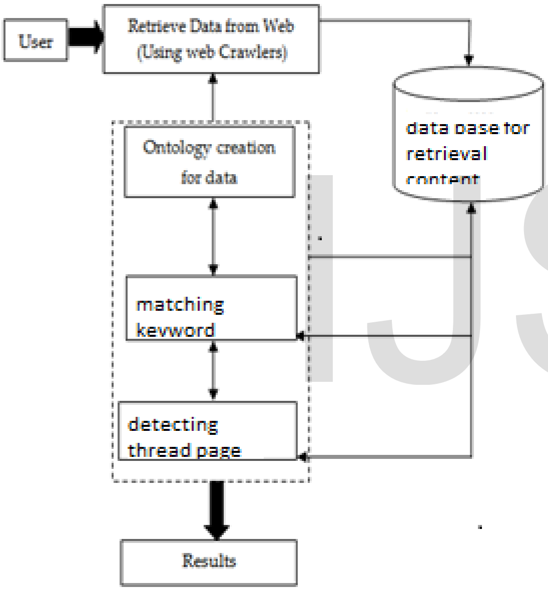

User
Retrieve Data from Web (Using web Crawlers)
data base for retrieval content
Ontology creation for data
matching keyword
detecting thread page
Results

Figure 1 Systematic Data Flow Diagram

4. DATA PREPROCESSING

Data Preprocessing is a Computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. In this process we use web crawlers to retrieve online data from web. A Web crawler is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing. A Web crawler starts with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the crawl frontier. URLs from the frontier are recursively visited according to a set of policies. Here, we use crawler to search the keyword for relevant topics of URLs from particular domain and it provide the interrelationship between concepts.

4 Support Vector Machine

Support vector square measure used for supervised learning in machine learning related to learning algorithms that analyze knowledge and acknowledge patterns used for classification and regression analysis. Given a group of coaching examples, every marked as happiness to one of two classes, Associate in Nursing SVM coaching formula builds a model that assigns new examples in to 1 class or alternative. Associate in Nursing SVM model is that the illustration of the examples as points in house, mapped so example of the separate classes square measure divided by a transparent gap that as wide as possible. New examples square measure mapped in to it some house and foretold to belong to a class based mostly on the that facet they fall on.

Mapping the high dimensional data in linear kernel support vector machine is not possible because of large dataset so RBF(Radial Basis Function) is used to measure the similarities between the two examples. It is also called as Gaussian kernel. The Gaussian kernel should be calculated by

$$K(\mathbf{x}_i,\mathbf{x}_j)=\exp(-\frac{\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2}{2\sigma^2})$$

Choice of kernel: Gaussian or polynomial kernel is default; if ineffective, more elaborate kernels are needed domain experts can give assistance in formulating appropriate similarity measures

Choice of kernel parameters

- e.g. σ in Gaussian kernel
- σ is the distance between closest points with different classifications

5. MULTI KEYWORD WEB CRAWLING

The existing system of this work deals with the removal of duplicate and unwanted pages from the forum crawling and the classifier support vector machine correctly classifies the URLs to achieve the highest performance for effectiveness and coverage. Compared to variety of crawlers it achieves the better performance and it also works with the question and answer forum sites and this system proposes the method by using the ontology. By using the multi keyword search under the specific domain to provide a better achievement.

5.1 Creating Ontology

Ontology is the philosophical study of nature of being, becoming, existence or reality basic categories of being and their relations. It deals with questions what entities should exist and how much entities can be grouped related to this hierarchy and sub divided according to similarities and differences. If that keyword matches to some features like super script, subscript, some related matching patterns then identify the correct page to retrieve.

We have already created dataset. That contains information about the different forum pages. We want to create ontology for every data by using following steps. Organizing and Scoping. The organizing and scoping activity establishes the purpose, viewpoint, and information for the ontology development project, and assigns roles to the team members. In the paper the ontology design of information about particular topic,inter related keywords are taken into account. During data collection, raw data needed for ontology development is acquired. Data analysis involves analyzing the data to facilitate ontology extraction. The initial ontology development activity develops a preliminary ontology from the data gathered. Ontology Refinement and Validation is done at the final stage. Here validation and comparison of final titles of the articles are considered .The ontology is refined and validated the ontology to complete the development process.

5.2 Protégé Tool

Protégé is a free, open source ontology editor and framework for the collection of knowledge and it can be exported in to variety of forms like OWL and XML Schema, it is flexible for rapid prototyping and application development.OWL which contains the description of classes, properties and their instances etc. Based on that the relevant information should be collected by using protégé to develop an ontology and this should accessed by using the multi keyword crawling.

5.3 System Overview

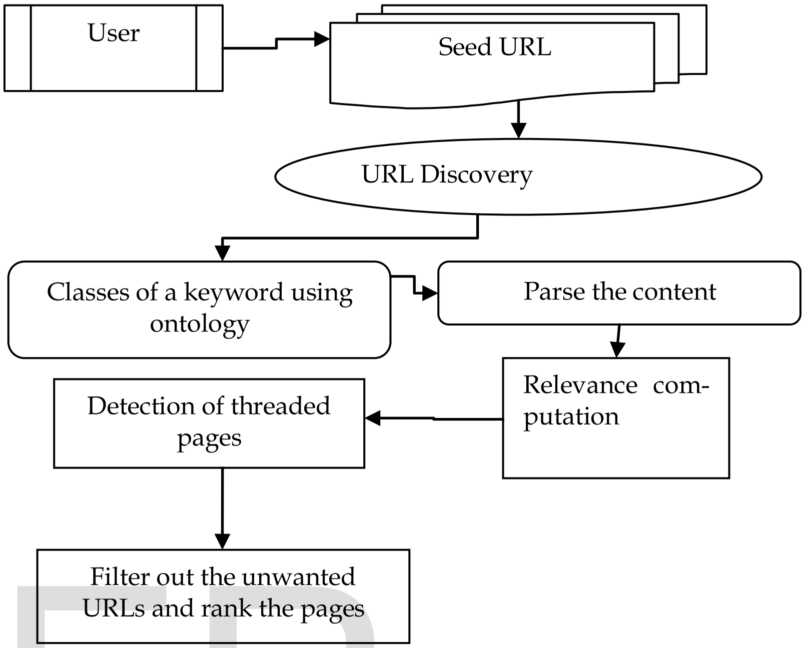

User
Seed URL
URL Discovery
Classes of a keyword using ontology
Parse the content
Relevance computation
Detection of threaded pages
Filter out the unwanted URLs and rank the pages

Figure 2: Multi Keyword Web Crawling

User seeds the URL to identify the entry URL discovery by using the entry URL discovery algorithm. The keyword matches with their class label for the relevant content to retrieve by using hierarchical structure. Parse the content and performing the relevance computation for identifying the relevant pages to match with the keyword to another class label to retrieve the data and remove the irrelevant URLs. And rank the pages for retrieving the highest rank page match with keyword for user query.

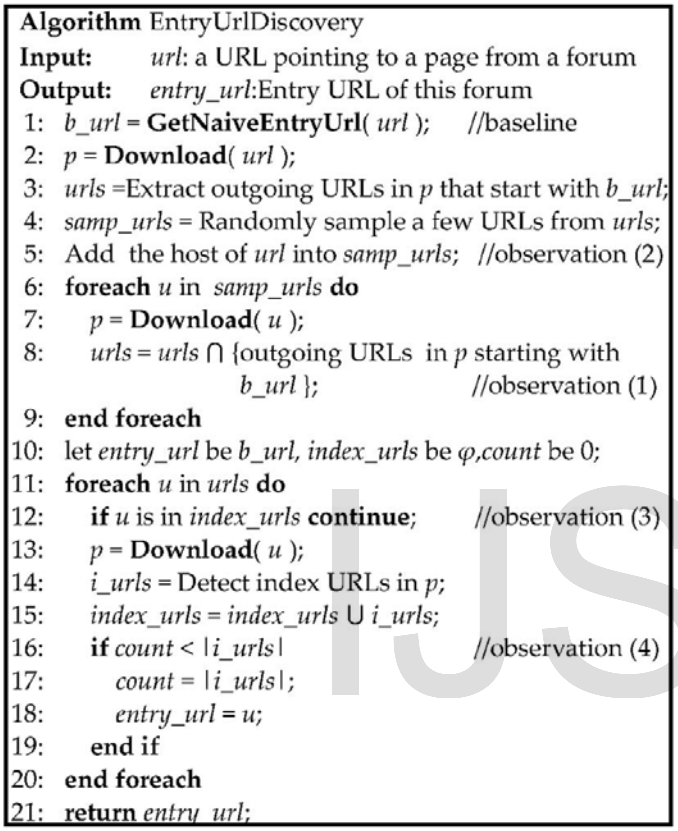
5.4 Entry URL Discovery Algorithm

Forum entry URL discovery is not a trivial task since entry URLs vary from forums to forums. To demonstrate this, we developed a heuristic rule to find entry URL as a baseline. The heuristic baseline tries to find the following keywords ending with “/” in a URL: forum, board, community, bbs, and discus. If a keyword is found, the path from the URL host to this keyword is extracted as its entry URLif not, the URL host is extracted as its entry URL.

1.Almost every page in a forum site contains a link tolead users back to its entry page. Note that thesepages are from a forum site.Aforum site might not be the site hosting this forum. For example, http://www. englishforums.com/English/ is a forum site but http://www.englishforums.com/ is not a forum site.

2. The home page of the site hosting a forum must contain the entry URL of this forum.

3. If a URL is detected as an index URL, it should not be an entry URL.



```
Algorithm EntryUrlDiscovery
Input:   url: a URL pointing to a page from a forum
Output:  entry_url:Entry URL of this forum
b_url = GetNaiveEntryUrl( url );      //baseline
p = Download( url );
urls =Extract outgoing URLs in p that start with b_url;
samp_urls = Randomly sample a few URLs from urls;
Add  the host of url into samp_urls;  //observation (2)
foreach u in  samp_urls do
    p = Download( u );
    urls = urls ∩ {outgoing URLs  in p starting with
                b_url };                //observation (1)
end foreach
let entry_url be b_url, index_urls be φ,count be 0;
foreach u in urls do
    if u is in index_urls continue;     //observation (3)
    p = Download( u );
    i_urls = Detect index URLs in p;
    index_urls = index_urls ∪ i_urls;
    if count < |i_urls|                 //observation (4)
        count = |i_urls|;
        entry_url = u;
    end if
end foreach
return entry_url;
```

6 Conclusion

Thus the paper gives us good results because of getting information from the all related topics and hence matching takes place between collections of dataset.Keywords are match with correct data the efficiency of matching will get increased. In this project, Ontology helps the crawler to extract and aggregate information from a specific domain. Using this method, matching feature between the keywords to identify the specific data for a specific domain. In the existing system the data can be retrieved from all the crawlers. According to the study on existing methods in previous papers of literature the downloaded pages can be retrieved this can be classified by using linear kernel SVM. But in this proposed method RBF kernel SVM can be used to perform high dimensionality large text data set. Using this technique the data can be effectively retrieved based on the keyword matching.

References

[1] Blog, http://en.wikipedia.org/wiki/Blog, 2012.

[2]"ForumMatrix,"http://www.forummatrix.org/index.php,2012

[3] R. Cai, J.-M. Yang, W. Lai, Y. Wang, and L. Zhang, "iRobot: An Intelligent Crawler for Web Forums," Proc. 17th Int'l Conf. World Wide Web, pp. 447-456, 2008.

[4] A. Dasgupta, R. Kumar, and A. Sasturkar, "De-Duping URLs via Rewrite Rules," Proc.14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 186-194, 2008.

[5] C. Gao, L. Wang, C.-Y. Lin, and Y.-I. Song, "Finding Question-Answer Pairs from Online Forums," Proc. 31st Ann. Int'l ACMSIGIR Conf. Research and Development in Information Retrieval,pp. 467-474, 2008.

[6] M. Henzinger, "Finding Near-Duplicate Web Pages: A Large-Scale Evaluation of Algorithms," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 284-291,2006.

[7] H.S. Koppula, K.P. Leela, A. Agarwal, K.P. Chitrapura, S. Garg,and A. Sasturkar, "Learning URL Patterns for Webpage De-Duplication," Proc. Third ACM Conf. Web Search and Data Mining,pp. 381-390, 2010.

[8] K. Li, X.Q. Cheng, Y. Guo, and K. Zhang, "Crawling Dynamic Web Pages in WWW Forums," Computer Eng., vol. 33, no. 6,pp. 80-82, 2007.

[9] G.S. Manku, A. Jain, and A.D. Sarma, "Detecting Near-Duplicates for Web Crawling,"Proc. 16th Int'l Conf. World Wide Web, pp. 141-150, 2007.

[10] V.N. Vapnik, The Nature of Statistical Learning Theory. Springer

[11] M.L.A. Vidal, A.S. Silva, E.S. Moura, and J.M.B. Cavalcanti,"Structure-Driven Crawler Generation by Example," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 292-299, 2006.