Abstract—This paper proposes an auto-classification algorithm of Web pages using Data mining techniques. We consider the problem of discovering association rules between terms in a set of Web pages belonging to a category in a search engine database, and present an auto-classification algorithm for solving this problem that are fundamentally based on Apriori algorithm. The proposed technique has two phases. The first phase is a training phase where human experts determines the categories of different Web pages, and the supervised Data mining algorithm will combine these categories with appropriate weighted index terms according to the highest supported rules among the most frequent words. The second phase is the categorization phase where a web crawler will crawl through the World Wide Web to build a database categorized according to the result of the data mining approach. This database contains URLs and their categories.

Keywords—Information Processing on the Web, Data Mining, Document Classification.

I. INTRODUCTION

EXISTING search engines such as Google (www.google.com), Yahoo (www.yahoo.com) and MSN (www.msn.com) builds and stores huge keyword-based indices that help locate sets of Web pages that contain specific keywords. These often return a long list of search results, ranked by their relevancies to the given query. Web users have to go through the list and examine the titles and short snippets to identify their required results. This is a time consuming task when keywords belong to many categories. For example, “jaguar” can be categorized under animals, cars, music, education, cities and more. In [9], a query for “jaguar” animal found in the 10th, 11th, 32nd and 71st pages of Google results.

Based on these observations, we believe intelligence should be integrated with the Web search engine service to enhance the quality of Web searches. A possible solution to this problem is to classify Web pages and search results into different categories, and to enable users to identify their required category at a glance. [13] Showed that relevant documents tend to be more similar to each other. This work could also contribute to concept-based search engines. Applying a concept-based search engine described in [12] shall include synonyms, super and sub concepts that will perform a parallel search in all domains and that will obtain an even larger set of documents than the search for the keywords alone would return. Classified search engine selects a smaller set of the search engine database index to search in. Thus provide the basis for discovering more relevant documents.

II. AUTO CLASSIFICATION OF WEB PAGES – FORMAL NOTATION

Given a database of Web pages D, where each Web page t is a list of terms (appeared simultaneously), Auto-Classification Algorithm is the technique that deals with the discovery of all the rules that correlate the presence of one set of terms with a category.

The following is a formal statement of the problem:

Let I is a Lexicon with a set of terms, \( I = \{i_1; i_2; \ldots; i_m\} \) be a set of m distinct terms. Let \( d \) be a set of n Web pages belonging to category \( c \), where each page \( t \) includes a set of terms such that \( t \subseteq I \). Associated with each category a unique identifier, called CID. Associated with each Web page is a unique identifier, called its TID. We say that a Web page \( t \) contains \( X \), a set of some terms in I, if \( X \subseteq t \) and \( X \subseteq I \). Hence, the database D is kept normalized and each database record is in the form <CID, TID, X>. A Web page of k terms is called k-termset.

An association rule is an implication of the form \( X \rightarrow Y \), where \( X \subseteq I \), \( Y \subseteq I \), \( X \subseteq t \), \( Y \subseteq t \) and \( X \cap Y = \emptyset \); The support...
degree is the co-occurrence frequency of set of terms within the category. The greater the degree of support, the more correct the mapping relationship between category and set of terms.

The rule $X \rightarrow Y$ has support sup in the pages set $d$ if sup% of Web pages in $d$ contain $X \cup Y$. Support($X \rightarrow Y$) = freq($X$, $Y$)/n.

The degree of confidence reflects the proportional of Web pages that contain both $X$ and $Y$ in $d$, to those that contain $X$ only. The rule $X \rightarrow Y$ holds in the category $c$ with confidence $\text{conf}$ if conf% of pages in $c$ that contain $X$ also contain $Y$. Confidence($X \rightarrow Y$) = freq($X$, $Y$)/freq($X$).

Given a set of Web pages $d$ belonging to a category $c$, the problem of mining association rules is to generate all association rules that have support and confidence greater than the user-specified minimum support (called minsup) and minimum confidence (called minconf) respectively.

III. THE AUTO-CLASSIFICATION STAGES

The proposed Auto-Classification technique has two phases. The first phase is a training phase where human experts determines the categories of different Web pages, and the supervised Data mining algorithm will combine these categories with appropriate weighted index terms according to the highest supported rules among the most frequent words.

The second phase is the categorization phase where a web crawler will crawl through the World Wide Web to build a database categorized according to the result of the data mining approach. This database contains URLs and their categories.

The Auto-Classification technique that discovers the category in which a set of Web pages belong to involves many stages:

1. Data Collection

To complete the training environment of the system, we need samples training Web pages that sufficiently represent the category. Hence, we collect source data to build up the source database. The source data is a set of categorized and pre-classified training set of pages. Let $T = \{t_1; t_2; \ldots; t_d\}$ be a set of Web pages belonging to category $c$, such that $T \subseteq D$.

2. Data cleaning

In order to mine the web pages, we need to extract salient terms. Therefore, some basic web page pre-processing is needed. It analyzes the Web pages, identifies the different HTML sections, parses and extracts the text. Extracted strings are normalized for processing. Linguistically, normalization of a word goes through a process known as morphological analysis. It first strips out all suffix and prefixes, stop words, and special characters.

In order to apply Apriori algorithm [19], it should be assumed that the terms are kept in their lexicographic order in each Web page. This will originate a set of salient terms $X \subseteq t$ and $X \subset I$ ready for data mining. We use the notation $x[I], x[2], \ldots, x[k]$ where $x[I] < x[2] < \ldots < x[k]$ to represent a $X$-termset.

3. Data Enrichment

A simple technique for extracting relevant terms is counting their frequencies in a given set of preprocessed documents. In general this approach is based on the assumption that a frequent string in a set of domain-specific extracted terms indicates occurrence of a relevant terms.

$Freq(string)$ is the probability of the string appearance in the Web page $i$. $freq$ determines strings of high frequency, i.e. greater than or equal minfreq. The result is a set of 1-termset (called $L_1$) which represent the candidate salient terms, such that $L_1 \subseteq X$.

Then we count the support of $L_1$ candidates $x_{sup}$. This represent is the probability that the term co-exist with the category in the source database and determine the seed set of termset for generating the set of potentially set of terms co-occur within a category.

The above stages will provide us with a data source ready for the terms association extraction.

4. Association rules extraction

The association rule extraction algorithm discovers the termsets, which represent the different associated terms in a category, through performing multiple passes over the data. A number of algorithms has been developed to generate candidate sets such AIS [20] and SETM [14]. We used Apriori algorithms [19] in this work.

Pass 1, we start with a $L_1$ termsets found in the previous pass to generate new potentially termsets, (candidates), and count their actual support.

Pass $k$, we start with a $L_k$ termsets found to be large in the previous pass to generate new potentially candidates and count their actual support.

This process continues until no new candidates are found. This determines the correlation of the set of terms and its category. Algorithm 1 below gives the association rule extraction algorithm used during the Training phase. Algorithm 2 gives the classification algorithm used during the crawled page classification phase.

Algorithm 1: Training

// extracting 1-itemset (candidates) from all Web pages belonging to a category $c$
Forall categories $c$ in $C$ do begin
  Forall pages $t$ in $c$ do begin
    Forall terms $x$ in $t$ do
      $X = \{x \mid x.freq \geq \text{minfreq}\}$
  End
  $L_1 = \{x \mid x.support \geq \text{minsup}\}$
// association rules extraction algorithm to discover termset
For ($k = 2; L_{k-1} \neq \phi; k++$) do begin
  $C_k = \text{apriori-gen}(L_{k-1})$; // generating new candidates using Apriori algorithm [19]
  Forall pages $t$ in $d$ do begin
    $C_t = \text{subset}(C_k, t)$; // Candidates contained in $t$
For all candidates \( c \in C_t \) do \( c: \text{count}++ \);
End
\( L_k = \{ c \in C_k \mid c: \text{count} \geq \text{minsup} \} \)
End
Answer = \( \bigcup \lambda L_k \); // set of rules R
// Repeat the above for all categories
End

Algorithm 2: Classification

For all terms \( x \) in \( t \) do
\( L_1 = \{ x \mid x.\text{support} \geq \text{minsup} \} \)
// association rules extraction algorithm to discover termset
For ( \( k = 2; L_{k-1} \neq \phi; k++ \) ) do begin
\( C_k = \text{apriori-gen}(L_{k-1}) \); // generating new candidates using Apriori algorithm [19]
For all pages \( t \) in \( d \) do begin
\( C_t = \text{subset}(C_k, t) \); // Candidates contained in \( t \)
For all candidates \( c \in C_t \) do \( c: \text{count}++ \);
End
End
// set of termset with highest support
\( A = \{ L_i \mid \max(L_i.\text{support}) \} \); // the category of the Web page based on the i-termset
Answer = \( \text{Similarity}(A) \);

The similarity function takes as argument \( L_\alpha \) the set of \( i \) terms in \( D \), and the set of i-termset returned from Web page \( t \) with the highest support. It returns the category \( \text{CID} \) of the matching set of items with highest support. The function works as follows

\[
\text{SELECT} \ \text{CID} \\
\text{FROM} \ L_i, p, t, L_i, q \\
\text{WHERE} \ p.\text{term}_i = q.\text{term}_i, \ldots \ p.\text{term}_i = q.\text{term}_i, \text{and} \max(p.\text{support});
\]

IV. THE STRUCTURE OF THE SYSTEM

The system of the search service comprises four main components that are depicted on Fig. 1. The main components are
- Web crawler
- Preprocessor
- Indexer and categorizer module
- Searcher
- Ranker

Next we describe the task of each component and how they operate.

- Web Crawler, fetches Web pages for parsing and indexing. It follows the links in each page to visit new pages, and eventually visit all the Web pages in the Internet. The Web page Meta information are stored in a repository for future visits. Some pages can be manually submitted to the Crawler to visit.

- Preprocessor, performs the data cleaning and data enrichment functions

- Indexer and Categorizer Module, The indexer engine creates, maintains the index catalog that serves as the basis of answering search queries. The categorizer engine administrates and determines the category information to support the category-based search of the system. The categorizer component performs supervised machine learning, i.e. it learns the significant words/expressions of each category of the taxonomy by means of sample training documents.

- Searcher, it provides a mechanism for keyword-based or concept-based searching. It accepts the user’s query, analyzes it, and rewrites the query (if necessary) using the specified thesaurus. Our Intelligent search engine query is imposed against the page index related to the specified domain.

- Ranker, sorts the resulted links to be displayed to the user. The Ranker computes the score of each result in the hit list, highlights the summary, sorts and presents the hit list.

V. VALIDATION EXPERIMENT

In our experiment we used a sample of 2000 documents from each of 13 categories to train our automatic classifier. Then we evaluated the performance by testing it against a new set of 500 randomly selected documents from each of the categories (not including documents used in the training phase). For both the training and classification we stripped the pages of their html tags.
Our classifier was given no priori information about the semantic content of the 13 categories, simply 2000 examples documents from each. After training our classifier, we are able to automatically produce a list of key words, which are the most distinguishing terms for each category. In Fig. 2, accuracy of the search results by a search engine that uses Vector Space versus our proposed search engine is depicted. The proposed algorithm outperformed the vector space algorithm in all categories in terms of accuracy. More experimentation are being studied right now to be included in future publication.

VI. CONCLUSION

In this paper, a classification algorithm of Web pages into a set of categories using data mining techniques is presented. The proposed technique is based on analyzing relationships between a set of documents and the terms they contain by producing a set of rules relating the category of the document, its terms and their frequencies. After training our classifier, we are able to automatically produce a list of terms, which are the most distinguishing for each category.

REFERENCES