An Idiosyncratic Tool for Retrieving Legal Web Documents Using SSARC algorithm

V. Annapoorani
Department of Computer Applications
Paavai Engineering College
Namakkal Tamilnadu
annapoorani504@gmail.com

***Abstract*- The practice of law necessarily involves a significant amount of research. In fact, the budding lawyers spend much of their work and time researching for the perfect information. Law and order is a field too vast, too varied and too detailed for any budding lawyer to keep all of it. Furthermore, the law is a living thing, it tends to change over time. Thus, in order to answer client's legal questions, lawyers typically conduct research into the laws affecting their clients. One of the most challenging problems is to incorporate domain knowledge in order to retrieve more relevant information from a collection based on a query given by the user.**

Keywords- SSARC, Semantic, Optimization, Physical

1. INTRODUCTION

The web domain contains lot of information on huge variety of documents in non-order groups. In this 21st century, all the details instantly available at our fingertips with the advent of Information Technology (IT). The searching of required information from huge database is difficult A.Smeulders.(2004),. Various search retrieval methods failed to retails the correct relational documents based on the knowledge process. In traditional method, we were limited to see through bookcases at the public library moving carelessly through the books in hopes of searching information. After the World Wide Web (WWW) developed, users gets all the information instantly. On WWW, the displaying list includes a link to the web page, the page's title and important points from the web page with the key term highlighted. With this modern era, the increasing availability of documents in digital form creates opportunities and challenges for all community and Information Technology researchers Chinatsu Aone, S. W. (2005). In this WWW, legal community also increases the number of legal documents using internet. While digitized documents facilitate searching all documents that are related to the task at hand and including a large number of them are not an easy task.

2. RESEARCH METHODOLOGY

In this SSARC algorithm, one can collect the legal documents from the web and to make available the whole collection list of civil cases to the lawyers. In our clock designing, similar to the normal clock, alphabet is used instead of the numbers 1 to 12. This SSARC algorithm is used to clustering methods for collecting and sorting the legal documents from the web Blair, D, Maron, M, E.(2008). This collecting, sorting, and storing work done by circular linked list in data structure, each node in the list is an alphabet, very first node collecting the a starting name of the legal documents. Each node in this linked list only holds the index of the civil case for the purpose of saving space. Each index consists of the links which holds the whole detail of that particular case Granger, (1977).

The below figure 1 shows the proposed implementation of Spontaneous Sorting and Retrieving Clock (SSARC) by various stages. This produce efficient search to retrieval legal web document have optimize best resultant with lower complexity. The following are the steps to process the documents analysis

2.1. Preprocessing

In this stage, the method reads the web document dataset holds the data point initialization. From the data points retrieved, the method identifies the nontrivial terms and list of the unique attributes of key terms J.Carbonell (2008). Then for each data terms performs Tokenization, Stemming, Remove Stop Words, and the method

verifies the presence of all the dimensions. If any of the data points have been identified as incomplete, then it will be removed from the data set. The preprocessed data set will be used to perform clustering.

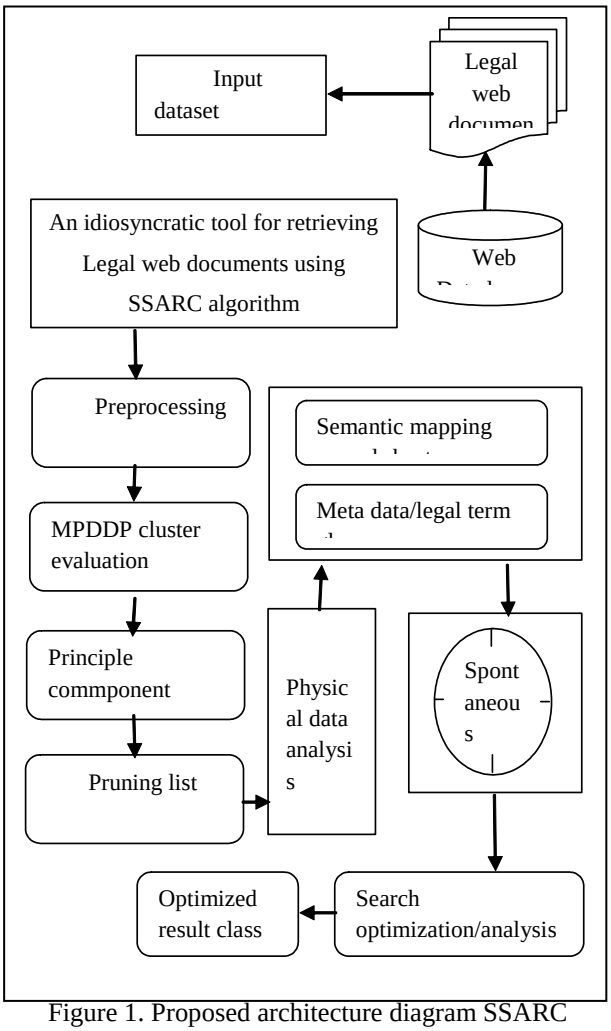

Figure 1. Proposed architecture diagram SSARC

2.2. MPDDP cluster evaluation

In this stage multi objective point of clustering based on Principal Direction Divisive Partitioning capable of partitioning a set of documents or other samples based on an embedding in a high dimensional Euclidean space, initially this performs number of cluster an point to Set Initial Cluster Center Randomly. Then Calculate the Distance Metrics to Select Highest Scatter Value cluster to split with analysis of PCA to form labeled pruning list. The method is unusual in that it is divisive document data grouping similarities, as opposed to agglomerative, and operates by repeatedly splitting clusters into smaller clusters. The documents are assembled in to a matrix which is very sparse.

2.3. Physical data analysis

In this stage the pruning list are actualized to take physical analysis of related search key terms, the method estimates multi point of relational terms. The method first identifies the source point of cluster points which holds the central data point of document terns. Then for each data term cluster, for each level of physical source is identified, the method computes the multi attribute data representation for each subspace identified.

2.4. Semantic mapping spread sheet

In this relational point of data analysis, the semantic sentence case relational terms are taken from the attribute mapping key term identification from cluster terms. This sematic patterns contains the real entity of

documentation holds the legal information. This evaluation is carried out through semantic lexicon induction to process the relational terms. The sentence contains the behavioral relational term factor originates the document reality in specialized mapping K.Collins-Thompson and J.Callan (2005). All the terms are relatively closed to the central relational key term. Using the semantic similarity based on singular vector decomposition to find the relative closeness of documents. The key terms are analyzed with extracted keyword related to ontological objectives. This analyses keyword relation between two terms based on the frequency level of patterns.

2.5. Spontaneous Sorting and Retrieving Clock (SSARC) optimization

The tremendous approach of web documents are handled using Spontaneous Sorting and Retrieving Clock in form of structuring the Meta data definition. The Meta data contains the information about the relational entity terms of documents similarity group. The documents are sorted based on the bootstrapping algorithm. Then the documents are in the form of an alphabetical order and stores the documents in clockwise structure algorithm Kimball. J. (1973). This constructs the indexing points of cluster evaluation to specify the legal documents from start point to end point based on PDDP cluster evaluation. Finally the optimization of documents retrieval is in legal collection based on the key term wording explicit from document set.

3. RESULTS AND DISCUSSION

The corpus contains 550 IP litigation case documents. A significant amount of noise was introduced in this data by this process. The corpus was preprocessed using an in-house tokenize and sentence boundary detector Salton, G. (1973). The sentence boundary was adapted to the pagination of this corpus, e.g., it introduces sentence breaks at two consecutive new line characters even if no punctuation mark exists. The resulting tokenized text was part-of-speech(POS) tagged using the standard POS tagger. Lastly, the corpus was annotated by an IP litigation expert, who followed strict annotation guidelines designed by a multi-disciplinary group of experts from both law and computer science. Table 1 summarizes the corpus statistics. This corpus was randomly split into a training partition(70%) and a testing partition(30%). This yielded a training corpus of 350 documents and a testing set of 150 documents.

Table: 1 Corpus Statistics

Documents	Sentences	Words	Claims	Claim Numbers	Claim Types	Patents	Laws
250	75,250	848,402	962	919	979	1892	633
300	80,250	1038,308	1254	1028	1024	2486	1345

Web data set was used for the evaluation of our proposed methodology. Preprocessing helped for identifying unique session and session id and removing noisy records. The proposed algorithms state the clear steps for preprocessing and session identification using the methods proposed in data preparation for mining World Wide Web browsing patterns.

3.1. Analysis of Time Complexity

Overall time taken to process the search retrieval document evaluation by retaining the result to optimization the cluster are as follows.

Figure 2 shows that the time complexity of clustering produced by different methods and it shows clearly that the proposed method has produced less time complexity than other methods.

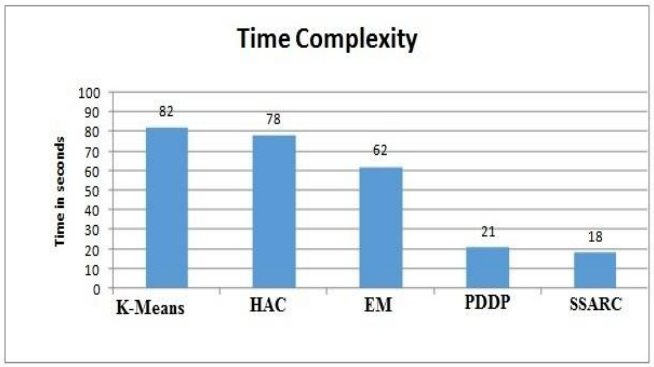

Figure 2. Time Consumption by various algorithms for generating the cluster

3.2. Overall performance.

Table: 2 Comparison result of overall performance

Algorithm	No. of Pages	Accuracy %
K-Means	2301	48.8
HAC	1560	89.8
EM	1504	96.4
PDDP	7089	98.97
SLIA	4356	99.37
SSARC	8282	99.92

From table 2, it is clear that the proposed algorithm produces more efficient and accurate indexing where the other methods produce less indexing accuracy. We conduct extensive experiments to validate our proposed approach. The results demonstrate the applicability of our approach and its capability of effectively identifying and categorizing mining services on the Internet web documents.

REFERENCES

A.Smeulders.(2004), H. a. Active learning using pre-clustering. ICML.

Blair, D, Maron, M, E.(2008). An evaluation of Retrieval Effectiveness for a Full- Text Document –Retrieval system,communications of the ACM .

Chinatsu Aone, S. W. (2005). Applying Machine Learning to Anaphora Resolution . In working Notes of the IJCAI-05 Workshop on New Approahes to Learning for Natural Language Processing,.

Granger, (1977). Foul-Up: A program that figures out meanings of words from context. In proceedings of the Fifth International Joint Conference on Artificial Intelligence.

J.Carbonell (2008). Paired sampling in density-sensitive active learning. ISAIM.

J.G.Carbonell. Subjective Understanding: Computer Models of Belief Systems. PhD thesis, Tech. Rept.150, Department of computer science, yale university, new haven.

K.Collins-Thompson and J.Callan (2005). Query expansion using random walk models. CIKM.

Kimball. J. (1973). Seven principles of surface structure parsing in natural language. Cognition.

Salton, G. (1970). Automatic Text analysis. Science .

Salton, G. (1973) Recent studies in automatic text analysis and document retrieval. Journal of the ACM,.