ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

An Innovative Aim for Collecting and Retrieving Documents from Web Domain Using SSARC (Spontaneous Sorting and Retrieving Clock) Algorithm

V.Annapoorani[1], Dr.A.Vijaya[2]

Assistant Professor, Department of MCA, Paavai Engineering College, Pachal, Namakkal, India[1]

Assistant Professor, Department of CSE, Govt Arts and Science College, Salem, India[2]

ABSTRACT: This paper presents an algorithm for generating and grouping documents from the web. In current years, due to the immense accessible of large document collections and the need to effective operate on them (for instance: navigate, analyze, query and summarize), there has been an increased emphasis on developing efficient and effective clustering algorithms for large document collections. In our novel algorithm collects all the documents from the web then it sorts the documents in an alphabetical order and stores the documents in clockwise structure algorithm which can easily retrieve the documents related to the user's query. This novel algorithm called as SSARC Algorithm, it is the expansion of "Spontaneous Sorting and Retrieving Clock" algorithm. We propose the overall architecture and depict two innovative algorithms which produce notable improvement over traditional clustering algorithms and form the basis for the query scrutinization and exploration of this algorithm.

KEYWORDS: Bootstrapping Algorithm, PDDP, SSARC, legal knowledge base, NLP, IE, AIR.

I. INTRODUCTION

Information Extraction is a subfield of NLP that is concerned with identifying predefined types of information from text. For instance, an IE system designed for a terrorism domain might extract the names of perpetrators, victims, physical targets, weapons, dates and locations of terrorist events or an IEs designed for a business domain might extract the names of companies, products, facilities and financial figures associated with business activities. NLP understanding is crucial for most IE tasks because the desired information can only be identified by recognizing conceptual role. The term "Conceptual Role" to refer to semantic relationships that are defined by the role that an item plays in context. Natural Language understanding systems that use conceptual knowledge structure typically rely on enormous amounts of manual knowledge engineering. While much of the work on conceptual knowledge structures has been addressed as initiate research in relating to the mental process involved in knowing, learning and understanding things of modeling and narrative understanding from a practical perspective. SSARC algorithm hoop around two magnitude concepts of scrutinizing and retrieving data. It constructs an index in a clockwise manner. In this structure, starting point collides with the ending point, so doubles up the speed of the same process done only in the forward process. In this algorithm collects any specific domain documents, but here collect the legal documents and storing in an alphabetical order by using bootstrapping algorithm. After that, it uses the PDDP algorithm to split the partition and clustering the similar and dissimilar legal documents. Automated retrieval from legal document collections were one of the most difficult test cases in legal fraternity. AIR (Automatic Information Retrieval) is primarily based upon concepts, not upon the explicit wording in the document texts. When manually examining legal texts, to realize important content, several aspects of the text need to be considered.

Our aim is to bring out an end – to – end legal information indexing system, which can give a solution to legal users for their day – to – day activities. The goal of SSARC algorithm combines with Bootstrapping and PDDP algorithm to collects the legal documents in an alphabetical order by using bootstrapping and after it collects the legal documents, it uses the PDDP to split the partition and clustering the similar and dissimilar legal documents. Automatic Indexing and Retrieval (AIR) concerns the selection of documents written in natural language from a database that are suitable for given information need. In a traditional IR system, a query composed of key terms is matched with the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

index terms of a document and upon matching that will be returned. In this thesis, the most specific problem is the Natural Language Understanding of document texts and user's performance. Retrieval of legal decisions is more complicated work and users usually want to retrieve an instance of the applications of some abstract concept. The lawyers to search similar cases to the one they have at already in hand. After a short definition of the architecture of SSARC in section 3, we narrate the clustering algorithm in section 4. The results obtained on a number of experiments using different methods to select sets of features from the documents show that partitioning clustering methods perform better than traditional distance based clustering. In section 5, we show how to use words obtained from clusters of documents to generate queries for related documents.

II. LITERATURE SURVEY

2.1 DISCO – Intelligent Help for Document Review – Jack O Neil.

This paper depicts a tool for helping lawyers and paralegal team is during document review in eDiscovery [3]. This tool combines a machine learning technology (CategoriX) and advanced multi-touch interface talents to not only oration the usual cost, time and precisions issues in document review, but to also speedup the work of the review teams by capitalizing as the intelligence of the reviewers and enabling conspire work.

2.2 CIRCUS – This paper developed to create case frame representation in response to sentence fragments, ungrammatical sentences with highly complicated syntactic structures are often directed without difficulty. CIRCUS does not involve complete dictionary coverage with respect to Part – Of – Speech recognition.

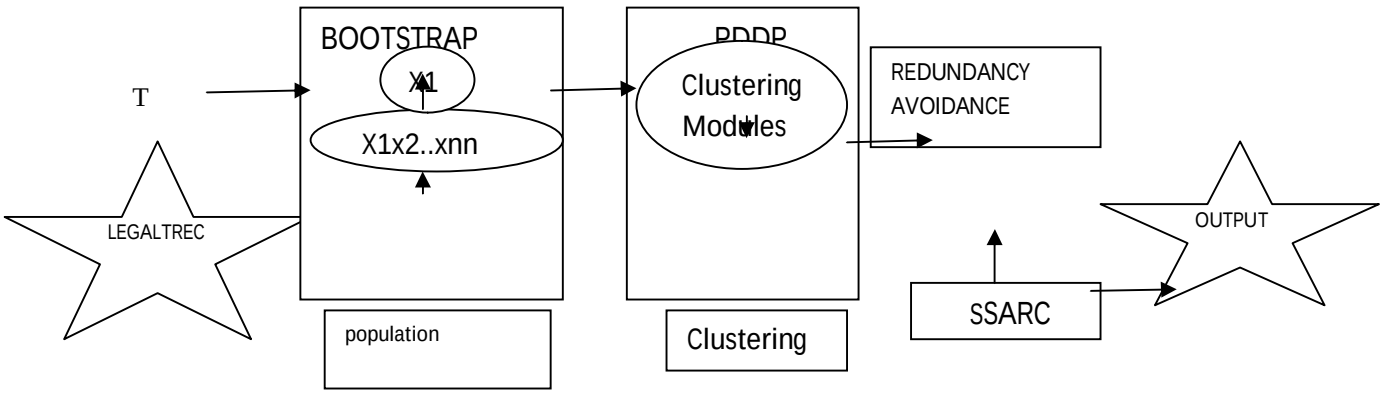
2.3 DISCERN SYSTEM – It developed a sub symbolic models is suitable approach to understanding mechanisms underlying natural language processing. It shows how individual connectionist models can be into a large, integrated system. This system faced with two major challenges are 1. How to implement connectionist control of high level processing strategies. 2. How to represent and learn abstractions.

2.4 A SEGMENT - BASED APPROACH TO CLUSTERING MULTI-TOPIC DOCUMENTS –

In this paper focus the problem of multi – topic document clustering by the power of natural configuration of documents in text segments, which bear one or more topic on their own. It provides a segment – based document clustering framework, which is designed to persuade a classification of documents starting from the recognition of cohesive groups of segment – based portions of the original documents. It provides substance of the consequence of our approach on different, large collections of multi – topic documents.

III. SSARC ARCHITECTURE

In this section, it illustrates our SSARC framework. It is exemplified as a framework rather than a singular specification because a number of the connected methods and implementation details can differ depending on the importance of the task and it provides a definite structure for building more specialized systems. A pictorial explanation of the framework as shown in fig1.



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

The proposal of this algorithm combines with Bootstrapping and PDDP algorithm. Its collects legal documents in an alphabetical order of sorting and retrieving the documents. This algorithm used the bootstrapping algorithm to collect the legal documents in an alphabetical order and after that uses the PDDP algorithm to split the partition and clustering the similar and dissimilar legal documents.

Automated retrieval from large document collections was one of the traditional applications of computer science. Today, text retrieval remains a specified component in legal information systems. However, the legal field provides one of the most difficult test cases. It is primarily based upon concepts, not upon the explicit wording test cases for automated text retrieval concerns the selection of documents written in natural language from a database that are suitable for a given information need. The goal of this project is to develop a system to create an alphabetic legal index for the needs of lawyers, judges and experts in the legal domain.

Our project aims at collecting legislative and judicial texts, as well as legal commentaries from federal, provincial and territorial jurisdictions in order to make primary sources of law accessible for free on the Internet. The large volume of legal information in electronic form creates a need for the creation and production of powerful computational tools in order to extract relevant information in a condensed form.

Thus the objectives of the present work from a technical perspective to:

1. This paper delivers a comprehensive survey on the method of creating an idiosyncratic tool.
2. An alphabetical sorting clock sorts and retrieves the legal judgments.
3. The goal of spontaneous sorting and retrieving clock is to make available the whole collection test of civil cases related documents to the lawyers.
4. Clusters the data collection and generates a set of distinctive word labels for each cluster of documents, all entirely autonomous.
5. All this processing by the tool occurs without input from a human user, except to specify the original document set.
6. In this clock, we sort the civil case labels in an alphabetical order.
7. Each node in the list is an alphabet very first node collecting ‘Aa’ is the starting name of the legal document.
8. Each node in this linked list only holds the index of the civil cases for the purpose of saving space.
9. Each index consists the links which holds the whole detail of that particular case.
10. Our sorting clock is capable of competitive performance in terms of speed, scalability and quality of class structure found.

In this study, combined bootstrapping and pddp algorithm to collect and sorting the legal documents. After finished this work, SSARC algorithm used to storing the legal documents in an alphabetical clockwise storage place. So, the users can take the desired documents based upon their queries from the web quickly.

3.1 BOOTSTRAPPING ALGORITHM

Natural language understanding requires both syntactic and semantic knowledge that contains semantic representation of all words, phrases and concepts in the language. It is unfortunate to expect a complete semantic knowledge based upon some restrictions of manual knowledge engineering. hence there have been two important efforts to build general - purpose semantic knowledge bases, WordNet(Miller 1990) and Cyc (Lenat, Prakash and Shepherd 1986). While general-purpose semantic information may be sufficient for some tasks, it is not to be sufficient for domain-specific applications. There are several advantages for creating a domain-specific lexicon. First, by definition, it contains the specialized finalize that is needed for in-depth understanding of the subject matter. Second, many complex problems in natural language can be understood by taking benefits of the limited domain. For instance, the word ‘monitor’ has several noun word senses, including one identifies to a computer screen and one that noted a lizard.

Many Natural Language Processing (NLP) systems depend on domain-specific lexicons, but these traditional are usually constructed by manual. It is taking a longtime and committed the errors to be lot. To address these problems, they have adopted a semi-automated approach to semantic lexicon construction.[2] This algorithm approach

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

is to begin with just a handful of seed words that are known to belong to a semantic class, and then leverage those seed words to find new words that also belong to the semantic class.

Conjunctions	Cows and Goats and Cats
Lists	Cows, Goats, Cats
Appositives	the horse, a black stallion
Compound Nouns	tuna fish; oak tree

Their affinity for semantically similar words can be exploited. The common idea of this algorithm is to begin with a small number of known category words and their identify other words that are collected near the known category words with unusual regularity.

3.2 PDDP (PRINCIPAL DIRECTION DIVISIVE PARTITION) ALGORITHM

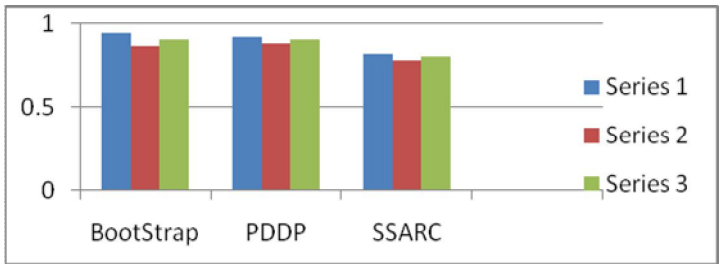
This algorithm is capable of competitive performance in terms of speed, scalability and quality of class structure found. The word 'Principal Direction' are used because the algorithm is based on the computation of the leading principal direction at each stage in the partitioning.[3] The key component in this algorithm that allows it to operate fast is a fast solver for this direction [3]. This principal direction is used to cut a cluster of documents repeatedly. The use of a distance or similarly measure is limited to deciding which cluster should be split next, but the similarity measure is not used to do the actual splitting. The word 'partitioning' to reflect the facts that place all the documents in a cluster, so that at every stage the clusters are disjoint and their union equals the entire set of documents.

This algorithm operates on a sample space of m samples in which each sample is an n-vector containing a numerical value. The algorithm proceeds by separating the entire set of documents into two partitions by using principal directions in a way. Each of the two partitions will be separated into two sub partitions using the name process recursively. The details of this algorithm are (1) what method is used to split a partition into two sub partitions. (2) In what order are the partitions selected to be split.

A partition of p documents is represented by an n x p matrix $M_P = (d_1,\ldots\ldots d_P)$ where each d, is an n-vector representing a document. The matrix M_p is a sub matrix of the original matrix M consisting of some selection of P columns of M, not necessarily the first P is the set, but we omit the extra subscripts for simplicity.

IV. EXPERIMENTAL RESULTS

Method	Precision	Recall	F-Measure
BootStrap	0.946	0.868	0.905
PDDP	0.924	0.886	0.904
SSARC	0.824	0.787	0.805



Method Name	Time Taken
BootStrap	695
PDDP	243
SSARC	115

In this paper, the objective of the assessment in this evaluation was two fold:
First fold is the clustering algorithm able to discover all the legal clusters in these documents.
Second fold is the clustering algorithm able to find the common legal issues in each of the reports.
Here, used Precision P and recall R to measure the performance for the objective, in which

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

-P is defined by the number of correctly identified clusters of a document (compared to the manually identified clusters) divided by the total number of clusters and -R is defined by the number of correctly identified clusters of a document divided by the total number of manually identified clusters of a document.

For the second objective, used precision P for evaluation, which is defined by the number of common legal issues identified among documents in all reports divided by the number of common legal issues manually identified among documents in all reports by experts.

Regarding these objectives, the precision and recall of the system on 10 reports across different document categories is shown in the figure 2. Overall achieved reasonably high precision, but the recall was quite low especially for case law documents. The main reason for this is the aggressive filtering, by adopting much higher thresholds, in the past processing of the system to achieve high precision.

V. CONCLUSION AND FUTURE REFERENCES

In this paper proposed a new method for sorting the branches of civil cases using SSARC algorithm in an alphabetical order. In this thesis, we have constructed a model sorting clock which sorts the branches of legal documents based on the titles spontaneously. The use of this method is not restricted to a single clock, it can be extended into concentric clocks in future. The most vital problem for information retrieval research now is to give us an efficient model for how large, operational retrieval systems work. If information retrieval research is successful in managing this transition, researchers can look forward to future work of a richness and complexity comparable to the recent history of database research with quick and detailed information.

REFERENCES

1. BoleyD. 1998. Experimental PDDP Software. http:// www.cs.umn.edu/~bolay/PDDP.html
2. Thilan, RillOff – "A Bootstrapping method for learning semantic lexicons using extraction pattern contexts" –- 2002
3. D.L. Boley, Principal Direction Divisive Partitioning, *Data Mining and Knowledge Discovery* **2**(4) (1998), 325-344.
4. D.L. Boley, M. Gini, R. Gross, S. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher and J. Moore, Document Categorization and Query Generation on the World Wide Web Using WebACE, *AI Review* **11** (2000), 365-391.
5. M. Shafiei and E. Milios. A statistical model for topic segmentation and clustering. *Lecture Notes in Computer Science*, 5032, 2008. Svm light. http://svmlight.joachims.org/, 2010.
6. A.Tagarelli and G. Karypis. A segment-based approach to clustering multi-topic documents. In *Proceedings of the Text Mining Workshop, SIAM Data Mining Conference*, 2008.
7. M.Utiyama and H. Isahara. A statistical model for domain-independent text segmentation. In *Proceedings of the ACL*, pages 499-506, 2001.
8. H. Kozima. Text segmentation based on similarity between words full text. In *Proc. of the ACL*, pages 286-288, 1993.
9. H. Kozima and T. Furugori. Similarity between words computed by spreading activation on an english dictionary. In *Proceedings of the ACL*, pages 232-239, 1993.
10. Berger, A. L., Della Pietra, S. A. & Della Pietra, V. J. (1996). A maximum entropy approach to natural language processing.Computational Linguistics, 22:39-71.
11. H.Nguyen and A.Smeulders. Active learning using pre-clustering. In ICML '04, pages 623-630, 2004.
12. Kimball. J. (1973). Seven principles of surface structure parsing in natural language. Cognition, 2. 15-47.
13. Cortes, C. & Vapnik, (2002). V.Support – Vector networks, machine learning, 20:273-297.
14. Blair, D.C. & Maron, M>E.(2008). An evaluation of Retrieval Effectiveness for a Full- Text Document –Retrieval system,communications of the ACM, 28(3), 280-299.
15. Salton, G.(1970). Automatic Text analysis. Science, 168, 3929, 335 – 343.
16. Salton, G. (1973). Recent studies in automatic text analysis and document retrieval. Journal of the ACM, 20(2), 258-278.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

BIOGRAPHY

V.Annapoorani

Annapoorani Venkatesan is pursuing Ph.D at Periyar University in the Master of Computer Applications. Her academic record is laden with First class throughout. She is currently an Asst.Professor in Department of MCA, Paavai Institutions, Namakkal, TamilNadu, India. She has a number of research publications to her credit in reputed National and International Conferences in the area of Data Mining and Human and Computer Interaction.

Dr. A. Vijaya

Dr.A.VIJAYA KATHIRAVAN is working as an Assistant Professor in Computer Applications in PG and Research Department of Computer Science, Govt. Arts College (Autonomous), Salem-07, TamilNadu, INDIA. She received her M.Phil. in Computer Science from Bharathiar University, Coimbatore and she awarded her doctoral degree in Computer Applications from University of Madras, Chennai. She has published 6 Books, 3 papers in National Journal, 30 papers in International Journal, 35 Papers in National Conference Proceedings, 38 Papers in International Conference Proceedings and a total of 112 publications. Her research interests include data structures and algorithms, data/text/web mining, search engines, web communities, social network mining, machine learning, Natural Language Processing, Organizational leadership and human resource management.