International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-2, Issue-5, September 2014

Time Reduction Mechanism in Information Extraction Using Parse Tree Query Language

K Venkatesh, B Vijaya Bhaskar Reddy

Abstract **- Information extraction (IE) is the task of automatically extracting structured information from unstructured and semi-structured machine-readable document. In this paper, we propose a new paradigm for information extraction. In this extraction framework, intermediate output of each text processing component is stored so that only the improved component has to be deployed to the entire corpus. Extraction is then performed on both the previously processed data from the unchanged components as well as the updated data generated by the improved component. Performing such kind of incremental extraction can result in a tremendous reduction of processing time. To realize this new information extraction framework, we propose to choose database management systems over file-based storage systems to address the dynamic extraction needs. To demonstrate the feasibility of incremental extraction approach, experiments are performed to highlight two important aspects of an information extraction system: efficiency and quality of extraction results.**

Keywords **—Text mining, query languages, information storage and retrieval**

I. INTRODUCTION

It is unsurprising that every year extra than 5, 00,000 articles are accessible in the biomedical writing, with secure to 20 million production access being put away in the Medline database. Extricating in succession from such a substantial corpus of archives is extremely convoluted. So it is vital to accomplish the extraction of data via automaticity. Data Extraction is the movement of concentrating organized data from the shapeless data. Incremental data extraction system utilizes database association framework as a key part. Database administration framework gives the element extraction needs over the record based storage room frameworks. Content preparing segments of named substance distinguishment and parsers send for the whole content corpus. The transitional yield of every content transforming constituent is put away in the social database proposed as Parse Tree Database (PTDB). Database question which is utilized to recover the data from the PTDB is as Parse Tree Query Language (PTQL). In the event that the recovery objective is altered or a module is effective than the relating module then just the dependable module is spread for the whole content corpus and the transformed information is possessed into the PTDB. At that point recovery of data is performed on the information that is included instead of whole content. Dissimilar to the record based pipeline loom, incremental in succession extraction skeleton weaving machine saves the transitional transformed data of every part; this keeps away from the prerequisite of reprocessing the complete content corpus. Keeping away from such reprocessing of data is most critical for data extractions on the grounds that it diminishes the extraction time greatly.

II. SYSTEM

This original extraction frame work consists of two segments. They are: Initial Phase used to giving out the text, Extraction Phase used to achieve the extraction.

Initial Phase: Text mainframe is responsible to execute a one-time parse, entity detection, and tagging on the entire corpus based on the current knowledge. This processed text is hoard in a relational database, call parse tree database (PTDB).

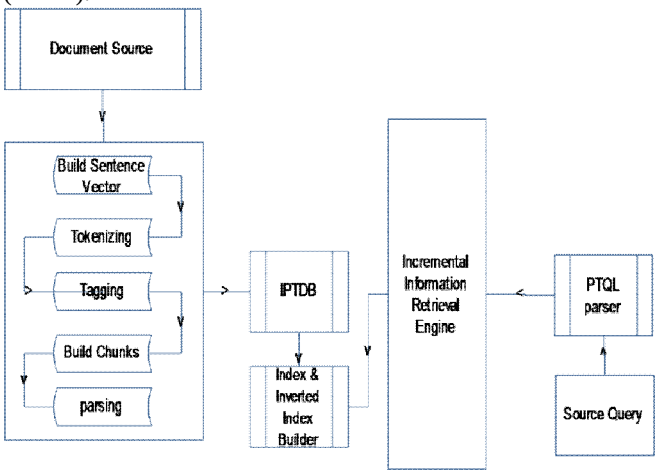

Figure 1. System Architecture

Extraction Phase: Extraction is then accomplished by PTQL. PTQL Query Evaluator transforms the PTQL query into keyword-based uncertainty and SQL queries. These are assessing by using the RDBMS and IR engine. Inverted index is comprehensive from the index designer to speed up the query assessment. This has been done by indexing the verdict according to the words and the equivalent entities. PTQL queries are produce using two modes of process. They are: training set determined query generation and pseudo significance feedback driven query production.Query Language for Information Extraction. Informationextraction is expressed as queries on the parse tree database.As query languages such as XPath and XQuery are notsuitable for extracting linguistic patterns [6], we designed andimplemented a query language called parse tree querylanguage, which allows a user to define extraction patterns ongrammatical structures such as

constituent trees and linkages.Since extraction is specified as queries, a user no longer needsto write and run special purpose programs for each specificextraction goal.

Automated Query Generation Learning the query languageand manually writing extraction queries could still be a timeconsumingand labor-intensive process. Moreover, such an adhoc approach is likely to cause unsatisfactory extractionquality. To further reduce a user's effort to perform information extraction, we design two algorithms toautomatically generate extraction queries, in the presenceand in the absence of training data, respectively.

Information Extraction IE has been an active research areathat seeks techniques to uncover information from a largecollection of text. Examples of common IE tasks include theidentification of entities (such as protein names), extractionof relationships between entities (such as interactionsbetween a pair of proteins) and extraction of entity attributes(such as co reference resolution that identifies variants ofmentions corresponding to the same entity) from text. Theexamples and experiments used in our paper involve the useof grammatical structures for relationship extraction. Co-occurrences of entities are a typical method in relationshipextraction, but often lead to imprecise results. Consider thatour goal is to extract relations between drug and proteinsfrom the following sentence: Quetiapine is metabolized byCYP3A4 and sertindole by CYP2D6. (PMID: 10422890)By utilizing our grammatical knowledge, a human readercan observe that hCYP3A4, metabolize, quetiapinei andhCYP2D6, metabolize, sertindole are the only correcttriplet relations for the above sentence. However, if weconsider co-occurrences of entities as a criteria to extractrelationships, incorrect relationships such as hCYP3A4,metabolize, sertindole and hCYP2D6, metabolize,quetiapinei would also be extracted from the abovesentence.

III. SYSTEM EVALUTION

Sentence Splitting: In the first module the documentscontain sentences. The sentences are in the unstructuredmanner. The module converts sentences to structuredsentences with index. This process is applied on the existingcorpus.

Word Indexing: In this module each sentence of adocument is made up with different words.

Example: S1= {w_1, w_2, w_3…….w_n}the module splits all the indexed sentences by words.

Word Tagging: In this module, the words will be presented inthe document in different forms such as present, past; futureetc…The words has to be n-grammed to find out the possibleequivalence of root words. The root words can be groupedtogether (or) clustered for special group of interests.Example: {"cricket", "football"} can be grouped together tospecial interests called "sports" category. Identifying group ofwords of similar category can have relationship. Building therelational words together is called word-net.

Parse Tree Database (PTDB) Construction: The word-net isa semantic relational network. The word-net is store in thedatabase as PTDB. The module provides an interface to theuser to search the PTDB of the corpus. The user's query willbe in the form of natural language (or) can be with stop words.

Execution Phase:

- The module provides as efficient way to query the PTDB
- The module provides an interface to the user to search the PTDB of the corpus.
- The user's query will be in the form of natural language (or) can be with stop words.

User's Query Preprocessing: In this module, user's query hasto be preprocessed against stop words elimination. The querywords have to be n-grammed for possible root words.

Query Word Tagging (PTQL): In this module, all the n-grammedwords may not be the root words. Find out the possible root words for each query word. Find thesemantically words for each word of query root word. Find theappropriateTag with their relevancies (or) Frequencies.

IV. PARSE TREE DATABASE (PTDB) SCHEMA

The parse tree database schema is illustrated in Figure1. Each tuple in the Constituents table corresponds to anode in the constituent tree of a sentence generated bythe Link Grammar parser. Each node is assigned withan unique CID in the Constituent table, and the fieldParentID indicates the CID of the parent node. Thefields Left and Right are in a way such that hierarchicalrelations between nodes can be easily identified,such as whether a node is a descendant of another node.Such labeling schemes are describe in Labeling Schema. Thefield WordOrder denotes the order of a word appearingin a sentence, while the attribute Sent_CID indicatesthe CID of the originating sentence. The table Linkageis used to store the links of the linkages of sentences.The attribute Link_Type indicates the types of the linksbetween pairs of nodes, which are stored as foreignkeys FROM_CID and TO_CID that refer to the field CIDin the Constituents table. The table Bioentitiesstores the entities that are recognized by the namedentity recognizers. Each of the entities has an uniqueBID and a foreign key that corresponds to a node in theConstituents table through the identifier CID. Withthe identifier BID, each entity is allowed to be assignedwith multiple entity types. The attributes StartByteand EndByte indicate the start and end bytes of anentity in its originating sentence.

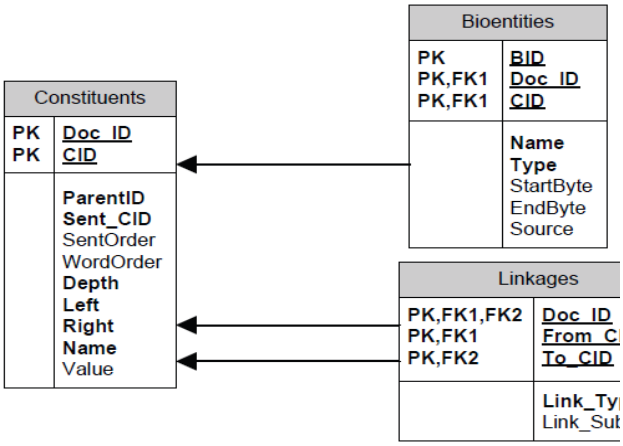

Figure 2. Database schema of the parse tree database(PTDB)

The tables Constituents and Linkages storethe constituent trees and linkages of the sentences thatare produced by the Link Grammar parser. The tableBio entities is for storing the entities recognized byvarious named entity recognizers.

LABELING SCHEME

Our labeling scheme follows the interval-based labelingscheme of the LPath language to capture the hierarchicalstructure and the horizontal relationships between nodesas well as the links between leaf nodes. Algorithm 1shows an algorithm to label a parse tree for a documentthat takes the root of the parse tree as input and returnsas output the tree with labeled nodes.

Algorithm 1 Labeling scheme
1: **if** tree node q = *null* **then**
2: Return;
3: **end if**
4: $q.id = id$, $q.d = depth$, $id = id+1$; {Set the id and thedepth for q and increment id}
5: **if** *parent* (n) _= *null* **then**
6: $q.pid$ = *parent* (n).id {Set q's parent id}
7: **else**
8: $q.pid = 0$
9: **end if**
10: **if** q is a leaf **then**
11: $q.l = left$, $q.r = left + 1$, $left = left + 1$
12: **else**
13: label q's children
14: **end if**
15: $depth = depth + 1$;
16: **for** each child p of q **do**
17: LABELNODE (p)
18: **if** p is the left most child of q **then**
19: $q.l = p.l$
20: **end if**
21: **if** p is the right most child of q **then**
22: $q.r = p.r$
23: **end if**
24: $depth = depth - 1$ {Decrement $depth$}
25: **end for**

V. MODULES

1) Data Arrival: Tuples arriving from each relation are initially stored in memory and processed

2) Cleaning Policy: Tracing relation between tuples, trace tuples designed to be in main memory

3) Managing statistics: Maintaining statistics for conceptual tuple region, update statistics during tuple arrival.

4) Reactive Phase: Performing joins between previously flushed data from both relations kept in discs

VI. RELATED WORKS

EUGENE Agichtein. LUIS Gravano[2]. We create a programmed question based method to recover archive valuable for the extraction of client characterized connection from huge content Database[9]. Assessment of PTQL inquiry includes the utilization of IR motor and additionally RDBMS. IR motor in inquiry is to choose sentence focused around lexical gimmick characterized in PTQL questions by RDBMS. Separating component depicted that select possibly important record for extraction. This separating still obliges filtering the complete database to consider each archive. We address the adaptability of data extraction framework in a principled and general way. Our methodology consequently finds the normal for archive that are helpful for extraction of target connection. Running data extraction framework over archive, we apply machine learning and data recovery procedure to inquiry that match extra helpful document[6]. Lawernce Hunter. Here we cover the outline, execution and a few assessment of Opendmap metaphysics determined, incorporated idea dissection framework. Opendmap data extraction frameworks were created for concentrating protein transport attestation and delivered quality is communicated in cell sort. Opendmap extricating protein communication forecast from full messages of biomedical examination articles. The yield of data extraction developed from component of a philosophy. The consequence of this exertion and give extra gimmick in frameworks that coordinate numerous hotspots for data extraction[8]. M. Banko, M.j. Cafarella, and Et al.[3] presents Open IE (OIE), another extraction standard where the framework makes a solitary information driven disregard its corpus and concentrates an expansive set of social tuples without obliging any human info. This paper additionally presents TEXTRUNNER, a completely executed, exceedingly versatile OIE framework where the tuples are doled out a likelihood and recorded to backing effective extraction and investigation by means of client inquiries. Content RUNNER performs the different procedures: initial one is self-directed learner that obliges no hand labeled information. The Second one is the Single-pass extractor. This Extractor makes a solitary ignore the whole corpus to concentrate tuples for all conceivable relations. Third is the Redundancy-based accessor, this Accessor appoints likelihood to each one held tuple focused around a

probabilistic model of excess in content. The dialect contains three expressive peculiarities, which are vital for semantic inquiry, in particular quick priority, Subtree perusing, and edge arrangement. Quick priority these primary level routes are intransitive, and none are backed by Xpath (in spite of the fact that their terminations are upheld). Subtree Scoping acquaints props into the dialect with grant degrees to be communicated. These will constrain all route to be obliged to a Subtree. In Edge Alignment Linguistic questions need to allude to hubs at the left or right edge of the subtree established at a defined hub.

VII. PROPOSED WORK

In the Initial Phase, an one-time parse, element distinguishment, and labeling (distinguishing every entrance as fitting in with a class of enthusiasm) in general corpus focused around the current earning is performed. The produced syntactic parse trees and semantic substance labeling of the transformed content is put away in a social database, called parse tree database (PTDB)[13]. Extraction Phase, Extraction is then attained by issuing database questions to PTDB. To express extraction designs, we composed and actualized an inquiry dialect called parse tree question dialect (PTQL) that is suitable for bland extraction. Note that in the occasion of a change to the extraction objectives (e.g., the client gets to be intrigued by new sorts of relations between entities)[7] or a change to an extraction module , the mindful module is sent for the whole content corpus and the prepared information are populated into the PTDB. Inquiries are issued to recognize the sentences with recently perceived notice. At that point extraction might be performed just on such influenced sentences as opposed to the whole corpus. Subsequently, to attain incremental extraction, this keeps away from the need to reprocess the whole accumulation of content not at all like the document based pipeline approaches. Utilizing database questions as opposed to composing individual exceptional reason programs, data extraction gets to be nonexclusive for assorted applications and gets to be simpler for the client. Notwithstanding, preparing information are not generally promptly accessible for specific connections because of the innate expense of making a preparation corpus. To gives the pseudo significance criticism driven approach that takes catchphrase based questions, and the PTQL inquiry generator then discovers regular syntactic examples among the top recovered sentences to produce PTQL inquiries from PTDB.

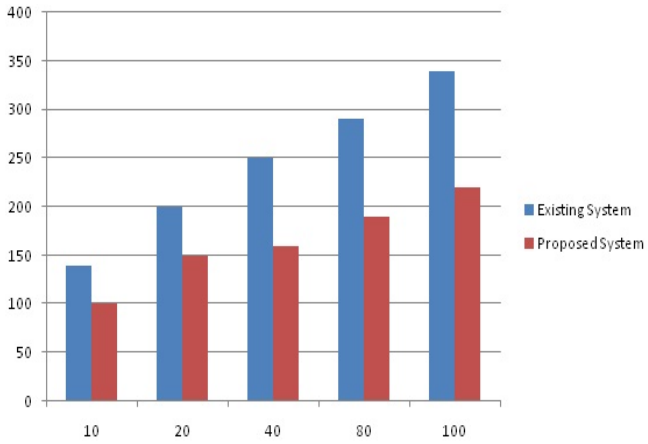

Figure 3: Time taken (Y-axis) for no of documents (x-axis)

VIII. CONCLUSION

Existing mining schemas don't give the capacity to deal with the middle of the road handled data. This prompts the preventable reprocessing of the whole content collection when the extraction objective is customized or enhanced, which could be computationally restrictive and lengthy one. To decrease this reprocessing time, the halfway transformed information is put away in the database as in unique schema. The database is in the presence of parse tree. To concentrate in arrangement from this parse tree the mining objective composed by the client in characteristic dialect content is changed into PTQL and after that extraction is execute on content corpus. This augmentation extraction weaving machine spares substantially more of an opportunity contrasted with performing mining by first transforming each one sentence each one in turn with semantic parsers and after that extra segments.

REFERENCES

[1] D. Ferrucci and A. Lally, “UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment,” Natural Language Eng., vol. 10, nos. 3/4, pp. 327-348, 2004.

[2] E. Agichtein and L. Gravano, “Snowball: Extracting Relations from Large Plain-Text Collections,” Proc. Fifth ACM Conf. Digital Libraries, pp. 85-94, 2000.

[3] M. Banko, M.J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, “Open Information Extraction from the Web,” Proc. Joint Conf. Artificial Intelligence (IJCAI), 2007.

[4] W. Baumgartner, Z. Lu, H. Johnson, J. Caporaso, J.Paquette, E. White, O. Medvedeva, K. Cohen, and Hunter, “An Integrated Approach to Concept Recognition in Biomedical Text,” Proc. Second Bio Creative Challenge, 2006.

[5] S. Bird, Y. Chen, S.B. Davidson, H. Lee, and Y. Zheng, Extending XPath to Support Linguistic Queries,” Proc. Workshop Programming Language Technologies for XML (PLAN-X), 2005.

[6] M. Cafarella, D. Downey, S. Soderland, and O. Etzioni, “Knowitnow: Fast, Scalable Information Extraction from the Web,” Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing (HLT ’05), pp. 563-570, 2005.

[7] J.T. Chang and R.B. Altman, “Extracting and Characterizing Gene-Drug Relationships from the Literature,” Pharmacogenetics, vol. 14, no. 9, pp. 577-586, Sept. 2004.

[8] F. Chen, A. Doan, J. Yang, and R. Ramakrishnan, “Efficient Information Extraction over Evolving Text Data,” Proc IEEE 24th Int’l Conf. Data Eng. (ICDE ’08), pp. 943-952, 2008.

[9] F. Chen, B. Gao, A. Doan, J. Yang, and R. Ramakrishnan, "Optimizing Complex Extraction Programs over Evolving Text Data," Proc 35th ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '09), pp. 321-334, 2009.
[10] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications," Proc. 40th Ann. Meeting of the ACL, 2002.

K. Venkatesh received his B.Tech degree from the department of Computer Science and Engineering from Sree Vidyanikethan Engineering College of Engineering, A.Rangampet, Tirupathi(Affiliated to JNTU Ananthapuramu). He is pursuing M.Tech from the department of Computer Science and Engineering in Shri Shirdi Sai Institute of Science and Engineering, Vadiyampeta, Ananthapuramu (Affiliated to JNTUAnanthapuramu). His current research interests include "Time Reduction Mechanism in Information Extraction Using PTQL".

Mr. B. Vijaya Bhaskar Reddy his currently working as an assistant professor in CSE Department at Shri Shirdi Sai Institute of Science and Engineering, Vadiyampeta, Ananthapuramu (Affiliated to JNTU Ananthapuramu). He received his B.sc computer science degree from CVLNR Degree College (Affiliated to SKU Ananthapuramu). He received M.Sc.-Computer Science from Sri Krishnadevaraya University College. He received his M.Tech from JNTUA College of Engineering. . (Affiliated to JNTU Ananthapuramu), His research interests in the area of Trust Negotiations in Peer to Peer Systems.