Data Abstraction and Visualization

Hrishikesh Wagh[1], Ganesh Pawar[2]
[1, 2] Department of Computer Engineering
[1, 2] Zeal College of Engineering and Research, Pune , Maharashtra, India

Abstract- *The data abstraction is most comprehensive topic and a need of time, although lot lots of abstraction and summarization tools exit there comes a need for a solution which could not just abstract and summarize data as per need but also help provide proper visualization to view data as need be.*

Since most of data available on internet or any other sources could be vast and would not always be possible for anyone to go through it, usually what we need is a solution which could narrow down the search to a key word and provide knowledge based keyword in time as per need. Still the data although summarize could be vast and hard to view, there require a solution to visualize data, to reconstruct it in a way that could be visualize and would make it easy to read.

Keywords- Data abstraction and Summarization, Data Mining, Natural language tool.

I. INTRODUCTION

This project is basically divided into three parts – Data preprocessing, Core processing and Visualization.

Data summarization is one of comprehensive topic many tools to summarize data already exit to which mostly does summarization though Auto summarization tools, summarized data is mostly about the core content of a text, although core knowledge can be found out by the processing through summarization it requires to have whole knowledge of text content and extract knowledge. This is where abstraction comes in picture, mostly scientific data would consist of factual data, and not imaginary. Scientific data is all about fact knowledge is mostly in present tense thus eliminating a need to have data in functional books. Take that under consideration most anomalies would be excluded for example when we consider and data from any sources like Internet, we are mostly concern with the facts and the knowledge with regard with actual key words. This would require to know actual key word in text and the extract sentence accordingly. This makes easy by abstracting sentence which are not of any concern with actual keywords. This is a way of abstraction. Considering scientific data allows to solve the problem of linguistic which might happen when it comes to non-scientific Data. Knowing that solution can be developed which concern itself with the scientific data, and does abstraction more conveniently.

When data is to be searched of Internet usually we go with keywords, which are mostly likely to be Know, comprehensive or comparative. Similar approach would allow us to abstract data. And provide with knowledge which concerns itself to Keyword approach. .

Knowing that summarization would not allow person to learn from the topic but might only provide suitable core knowledge which is yet to be explore in deep, while abstraction could abstract data and give similar content but would know guarantee the size and thus might not be suitable to view such enormous data, This Visualization comes into picture

Thus we will implement the system to search the text by keywords entered and provide not just summary but abstracted view of data with proper visualization tool which could make learning easy.

II. LITERATURE SURVEY

2.1 NLP: We point out some relevant issues that are related to the computing-with-words (CWW) paradigm and argue for an urgent need for a new, nontraditional look at the area, since the traditional approach has resulted in very valuable theoretical research results. However, there is no proper exposure and recognition in other areas to which CWW belongs and can really contribute, notably natural-language processing (NLP), in general, and natural-language understanding (NLU) and natural-language generation (NLG), in particular. First, we present crucial elements of CWW, in particular Zadeh's protoforms, and indicate their power and stress a need to develop new tools to handle more modalities. We argue that CWW also has a high implementation potential and present our approach to linguistic data(base) summaries, which is a very intuitive and human-consistent natural-language-based knowledge-discovery tool. Special emphasis is on the use of Zadeh's protoform (prototypical form) as a general form of a linguistic data summary. We present an extension of our interactive approach, which is based on fuzzy logic and fuzzy database queries, to implement such linguistic summaries. In the main part of the paper, we discuss a close

relation between linguistic summarization in the sense considered and some basic ideas and solutions in NLG, thus analyzing possible common elements and an opportunity to use developed tools, as well as some inherent differences and difficulties. Notably, we indicate a close relation of linguistic summaries that are considered to be some type of an extended template-based, and even a simple phrase-based, NLG system and emphasize a possibility to use software that is available in these areas. An important conclusion is also an urgent need to develop new proto forms, thus going beyond the classical ones of Zadeh. For illustration, we present an implementation for a sales database in a computer retailer, thereby showing the power of linguistic summaries, as well - - as an urgent need for new types of proto forms. Although we use linguistic summaries throughout, our discussion is also valid for CWW in general. We hope that this paper-which presents our personal view and perspective that result from our long-time involvement in both theoretical work in broadly perceived CWW and real-world implementations-will trigger a discussion and research efforts to help find a way out of a strange situation in which, on one hand, one can clearly see that CWW is related to words (language) and computing and, hence, should be part of broadly perceived mainstream computational linguistics, which lack tools to handle imprecision. These tools can be provided by CWW. Yet, CWW is practically unknown to these communities and is not mentioned or cited, and---reciprocally---even the top people in CWW do not refer to the results that are obtained in these areas. We hope that our paper, for the benefit of both the areas, will help bridge this gap that results from a wrong and dangerous fragmentation of break science.

2.1.1 A Preprocessing Framework Based on English Grammar, variation of Scientific and Nonscientific data

The entities in this system are: Uses laws on grammar in scientific data which is fact based, any data available in Studies text books or example like Wikipedia on Internet would be consider as factual data, fiction story books would not come under scientific data. Narrowing down the linguistic problem which arises due to different way or approach of writing, The actual system intend to find actual knowledge though key search taking under considering laws which comes with Scientific data. Such data would be factual and would not content any misplaced knowledge or data which would not be any used to a learner. This theory if put to test could extract the factual data, eliminating the need to deal with linguistic problems occurring with non-scientific function text data.

The whole preprocessing of this scientific data would happen with taking under consideration the rules of Grammar which are solid.

This approach eliminates the non-factual or non-scientific data from scientific data from which actual learning is to be done.

2.1.2 Linguistic barrier and problem:

2.1.3 Linguistic problem in data summarization :

In this paper, although data is abstracted problem with abstraction and summarization is mostly the same, linguistic problem comes with different ways of writing. Since text coming from different sources and from different authors it is highly unlikely to follow a similar approach or Structure and might vary considerably, which cause a major trouble while abstraction and summarization. This require for a Solution to have a consciousness which could identify a common structure and generate summary without wasting much time with implementing other rule.

2.2 Classes of Abstraction/ Summarization: we categorize class of Abstraction/ Summarization in two classes. Conceptual and Linguistic approach.

2.2.1 Conceptual class of Data Abstraction/ Summarization:

The author suggest Conceptual data abstartion taking under consideration the Key enterd for search. In which case following a close structure of text data can be extracting only knowing the realated Keywords without much concerning itself with having a need to know what an actual text stands for. This approach start with make Headlines, or Topic name its major keyword or a particlar paragraph as its major focus.

2.2.2 Linguistic class of Data Abstraction/ Summarization:

A approach suggest as a class by author requires to have a solution which would have a consciousness and can determine where a system should focus. This would require to build an system, literally an consciousness which would analysis structure of text content for example which category text fall under which could be Essay, Report, fiction or non-fiction. Author suggest being acquainted with such types could narrow down search to a level which not only provide focus but also save time or implementing another way which could reduce overhead to higher extend. .

Demonstrate via proposed measure and PCC can achieve better result for similarity measure than traditional PCC.

Using PCC measure, we can filter out some users' pairs that have more or less similarity to rating scores on the same items. Based on the assumption users with similar tastes on different types of products have higher probability to form a community and are likely to make associates to each other even if they don't know each other before.

III.SYSTEM ARCHITECTURE OVERVIEW

3.1 NLP (Natural Language Text Processing):

Natural Language toolkit or NLP (Natural Language processing) Serves a best tool for any processing on Natural Language, here Natural Language being English as in this System.

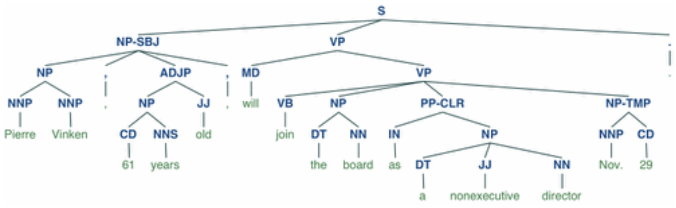

Fig. 1. Parser Tree

The NLP provides way to facilitate both class as describe in this paper, we intend to utilize the NLP way of processing to distinguish between class and provide way to abstract data allowing original plaintext to be summarized with accordance with normal processing.

3.2 Preprocessing:

Here is the first step which intend to differentiate text into scientific and non scientific this allowing only the factual scientifc data to be sored and eleminating other liguistic problems which might arrise in non scientific text.

3.3 Core Procssing:

Here an actual abstraction and summariztion is to be donw, in concern with keywords enterd by user and performing taking abstration/ summarization taking under consideration two class that is, conceptual and linguistic class as identified by author.

IV. METHODOLOGY

1. Data User: The data used here although could come from different sources would be flawlessly captured and provide transparence without much concern of preprocessing steps performs.
2. Preprocessing: Although important works in background provides a suitable way to extract factual data out from raw data which saved differently from original data used by Data owner.

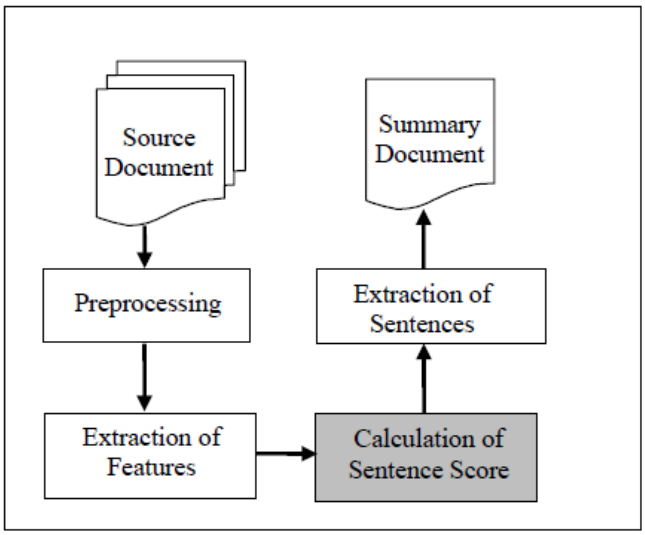

Fig. 3. System Architecture

Step 1: Data saved my User: The actual data saved by user is required to be categorized in scientific and non-scientific data.
Step 2: Selection Process : Allows user stored data to be selected for further core processing identifying a need to preprocess first .
Step 3: Enter Keyword: Allows user to enter keyword and search sentences and related sentence using algorithm based on conceptual and linguistic approaches.
Step 4: Visualization: Provide a tool with accurate view of abstracted/ summarized data. So that learner or reader could view easy and efficient.

V. CONCLUSION

This project establishes the importance of providing a tool which not only summarized text data but also provide abstracted data depending up user search keyword. Since summarization does not provide detail information there comes a need to abstract data, however the abstracted data although detailed could be enormous and would be hard to visualize, this having a need to have a separate visualization tool to view this data. The solution suggests work with scientific data.

In this report we suggest way to efficiently summarize or abstract text data. Taking under consideration the problems associated with linguistic and this enhancing ability to provide abstract data, which is required by a learner or reader.

There are still many problem associated with text processing and not all can be solved but this approach suggest to narrow down problem and focus on Scientific data.

ACKNOWLEDGEMENT

We would like to thank our Guide Prof. H. H. Patel for the support and guidance she gave us on every step of the project execution. We would also like to thank to, HOD, Prof. S. M. Sangve who gave us his valuable comments.

REFERENCES

[1] H. P. Luhn, "The Automatic Creation of Literature Abstracts" IBM Journal of Research and Development, vol. 2, pp.159-165. 1958..

[2] J. Kupiec. , J. Pedersen, and F. Chen, "A Trainable Document Summarizer" In Proceedings of the Eighteenth Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR), Seattle, WA, pp.68-73.1995.

[3] J. Kupiec. , J. Pedersen, and F. Chen, "A Trainable Document Summarizer" In Proceedings of the Eighteenth Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR), Seattle, WA, pp.68-73.1995..

[4] Amy J.C. Trappey, Charles V. Trappey, "An R&D knowledge management method for patent document summarization" Industrial Management & Data Systems, vol.108. pp.245-257. 2008..

[5] Divya S., Dr. P. C. Reghuraj Department of Computer Science and Engineering Govt. Engg. College Sreekrishnapuram "Eigenvector Based Approach for Sentence Ranking in News Summarization", International Journal of Computational Linguistics and Natural Language Processing Vol 3 Issue 3 March 2014.

[6] A.R.Kulkarni1, S.S.Apte2," An Automatic Text Summarization Using Lexical Cohesion And Correlation Of Sentences " , IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308.

[7] Divya S., Dr. P. C. Reghuraj Department of Computer Science and Engineering Govt. Engg. College Sreekrishnapuram "News Summarization Based on Sentence Clustering and Sentence Ranking".

[8] Divya S., Dr. P. C. Reghuraj Department of Computer Science and Engineering Govt. Engg. College Sreekrishnapuram "News Summarization Based on Sentence Clustering and Sentence Ranking".