A Novel Approach for Filtering Unrelated Data from Websites Using Natural Language Processing

Foram Joshi[1] Sandip Chotaliya[2]

[1] Student, Department of Computer Engineering, Noble Engineering College, Junagadh, Gujarat, India.
[2]Asst. Professor, Department of Electronics & Comm., Noble Engineering College, Junagadh, Gujarat, India.

ABSTRACT

Day by day review or opinion can be taken by number of websites. Either it's related to movie review or anything else. Every times it's not necessary that user post his/her opinion or review for particular subject only.so to filter out such unrelated comments or review we proposed a structure which is based on natural language processing. Our proposal first extract the comments or reviews from the particular site using web crawling concept and then it will processed on such data using natural language processing.so finally we will get the data which is only related to particular post.

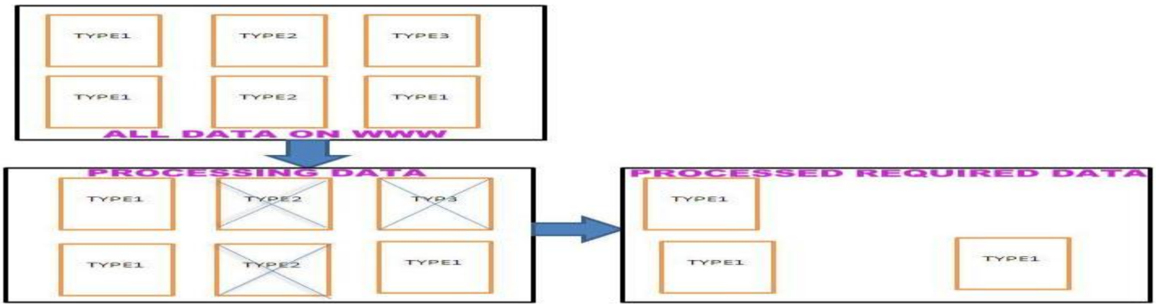
***Key Words**: Data Extraction, Natural Language Processing, Wrapper , Opinion, Web Crawling, seed, fetch*

I. INTRODUCTION

World is full of valuable data. In all that data to get our required or related data in a formatted way it is not easy task product listing, Business directories, Inventories etc data managing is very tedious.so that there are number of technique available and based on those technique number of soft wares are available to analyze the data. We are going to implement such intelligence technique among them by which we can easily manipulate the data [1].In this paper we are going to mash up three different concepts like data extraction, web crawling which is used for data extraction process and last natural language processing for processing that particular extracted data. Traditional approach for extracting data from web source is to write specialized programs, called Wrappers. what wrapper exactly do to identify data of interest and map them to some suitable format like relational database or XML.[3] In other words, the input to the system is a collection of sites, (e.g. different domains), while the output is a representation of the relevant information from the source sites, according to specific extraction criteria.[4] we can applied such technique for data extraction purpose to different types of text like newspaper articles, web pages, scientific articles, newsgroup messages, classified ads, medical notes etc.[2]

II. HOW EXTRACTION WORK?

First machine find numbers of data when we want to some specific type of data for extraction then it filter another data. Take a look in figure which gives basic idea. Figure represent that initially we want the data of type1 then what machine started processing all the data. And from that it gives our required data that is type1 data. This is what we can say wrapping the data. Now as we discuss that number of techniques are available for data extraction like natural language processing, language and grammars, machine learning, information retrieval, database and ontologies are there. [3] In those different technique we are going to implement natural language processing technique. During my work on this topic I found if we want to reliable data in simple way. Then we will go for natural language processing technique. Here one more diagram which give the flow of extracting the data. [Figure 2]



Figure[1]:Meaning of Data Extraction in pictorial format.

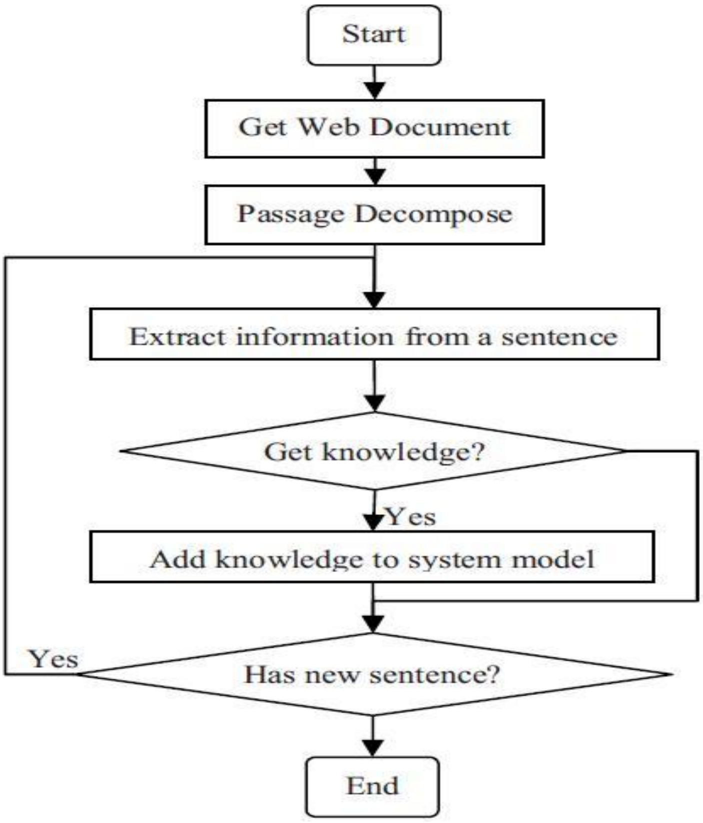


Figure [2]: Flow of extracting the data.

From Figure [2] we get basic idea that how data extraction can be done. This is basic idea of data extraction we are going to extract particular data from particular website so here we introduce one more concept that is web crawling concept.

III. HOW WEB CRAWLING CONCEPT WORK IN OUR PROPOSED STRUCTURE

- **Web crawling**
  Generally crawling is depending on string or parameter which are given to it as parameters which is called Seeds, they are added to a URL request queue. Crawler starts fetching pages from the request queue. And then parsed that downloaded page to find link tags that might contain other useful URLs.New URLs added to the crawler's request queue, or frontier. It will continue until no more new URLs or disk full.

- **Freshness**
  Web pages are constantly being added, deleted, and modified Web crawler must continually revisit pages it has already crawled to see if they have changed in order to maintain the freshness of the document collection stale copies no longer reflect the real contents of the web pages
- **Focused Crawling**
  Attempts to download only those pages that are about a particular topic used by vertical search applications. and rely on the fact that pages about a topic tend to have links to other pages on the same topic popular pages for a topic are typically used.
  Basically Crawler uses text classifier to decide whether a page is on topic.

IV. PROCESSING STEPS IN NATURAL LANGUAGE PROCESSING

Natural language processing is the automatic ability to understand text or audio speech data and extract valuable information from it.[2] the ultimate objective of natural language processing is to allow people to communicate with computers in much the same way they communicate with each other. More specifically, natural language processing facilitates access to a database or a knowledge base, provides a friendly user interface, facilitates language translation and conversion and increase user productivity by supporting English like input.[4] natural language processing is defined in vast area where it has been used either it would be main field like automatic summarization,coreference resolution, discourse analysis, machine translation, morphological segmentation, named entity recognition, natural language generation, natural language understating, optical character recognition etc or it may be used in sub fields like information retrieval, information extraction, speech processing etc.[5]

- Morphological Analysis: Individual words are analyzed into their components and non word tokens such as punctuation are separated from the words.
- Syntactic Analysis: Linear sequences of words are transformed into structures that show how the words relate to each other.
- Semantic Analysis: The structures created by the syntactic analyzer are assigned meanings.
- Discourse integration: The meaning of an individual sentence may depend on the sentences that precede it and may influence the meanings of the sentences that follow it.
- Pragmatic Analysis: The structure representing what was said is reinterpreted to determine what was actually meant.

V. PROPOSED STRUCTURE FOR OPINION EXTRACTION

Figure [3] gives clear idea that how number of input we take. And based on some pre decided rules and thesaurus we can categorized review.Actually what we are going to implement is first we extract solutions from website discussproblems.com which is based on sharing problems and giving solution for it. Extracting that data with the help of web crawling conception that we are processing on text with help of pre defined rules that based on which criteria particular comment or opinion is good, medium or any abuse thing. If there are N rules matching the same piece of text, we first rank rules preliminarily according to their own extracting accuracy [9].

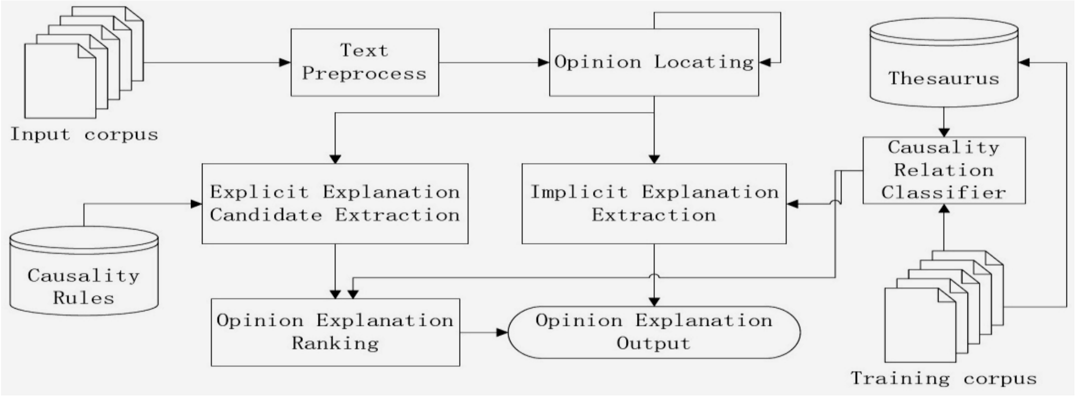

Input corpus
Text Preprocess
Opinion Locating
Thesaurus
Causality Relation Classifier
Explicit Explanation Candidate Extraction
Implicit Explanation Extraction
Causality Rules
Opinion Explanation Ranking
Opinion Explanation Output
Training corpus

Figure [3]: The process of our extraction method

And using Natural Language Processing various step we easily identify either posted comment is positive or negative. So we any unrelated posts are found then it will be filtered out.

VI. PARAMETER WORK ON

For checking particular sentence some parameters come into picture. With help of that parameters we easily manipulate for particular statement.

- **Precision**

$$Precision = \frac{\#\text{ correct answers}}{\#\text{ answers produced}}$$

- **Recall**

$$Recall = \frac{\#\text{ correct answers}}{\#\text{ total possible corrects}}$$

- F-measure

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

So, these are the basic parameter on which we can find the correctness of the particular statement.

VII. EXPERIMENTAL DATA

For Experimental purpose we developed one site that is DiscussProblems.com with the help of antique brains technology.DiscussProblems.com site is almost fulfill our requirement i.e what we want tht such site which is based on review or some what like it. This is same kind of site means here anonymous user post or share their problem with out giving his/her identity and other viewer give their better solution. So from there are number of categories are available for posting the problem and sharing their solution

VIII. Conclusion And Future Work

In this paper, we describe a novel approach for opinion extraction using Natural language processing and identify whether opinion is good ,bad or if it abuse type then it will automatically removed. Future work for this approach is its implementation for the various web sites which is based on review type or opinion based.

ACKNOWLEDGEMENTS

We really thankful to Antique brains Tech.,Rajkot to appreciate our work and giving required resources for practical purpose.

REFERENCES

[1] Mozenda Web Scraper - Web Data Extraction http://www.youtube.com/watch?v=gvWGSBRuZ5E

[2] Natural Language Processing http://en.wikipedia.org/wiki/Natural_language_processing

[3] Natural Language Processing 68 www.hit.ac.il/staff/leonidm/information-system/ch-68.html

[4] Natural Language Processing http://www.seogrep.com/natural-language-processing/

[5] Yuequn Li, Wenji Mao1, Daniel Zeng, Luwen Huangfu1 and Chunyang Liu A Brief Survey of Web Data Extraction Tools

[6] Mary D. Taffet Application of Natural Language Processing Techniques to Enhance Web-Based Retrieval of Genealogical Data

[7] PARAG M.JOSHI, SAM LIU. Web Document Text and Images Extraction using DOM Analysis and Natural Language Processing. To be published in the 9th ACM Symposium on Document Engineering, DocEng'09, Munich, and Germany. September 16-18, 2009

[8] Jagadish S KALLIMANI, Srinivasa , Information Extraction by an Abstractive Text Summarization for an Indian Regional Language

[9] Yuequn Li, Wenji Mao, Daniel Zeng, Luwen Huangfu1 and Chunyang Liu, Extracting Opinion Explanations from Chinese Online Reviews

[10] Foram Joshi Extracting Opinion From Web Sites Using Natural Language Processing, 9th National Confererence On Information, Knowledge & Research In Engineering, Technology & Sciences 2013
ISBN: 978-81-906220-9-7

[11] Foram Joshi A Comparative Approach For Data Extraction From Web Sites Using Natural Language Processing International Conference On Advanced Computer Science And Information Technology (ICACSIT)
GOA- ISBN: 978-81-925751-5-5