International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 5, May - 2014

A Paper on Approaches for Information Extraction from Unstructured Text

Piyush Gusani
CE department, NGI (Noble Engineering College)
Junagadh, India

Abstract- In today's computer world everything is converted into e world. Like , e-business,e –library-malls etc. That includes lots of e documents, information on the internet. Now the e documents is nothing but text, images etc.So from this available data to find the data of our interest is a complex task. We need information extraction systems to perform this task. This paper includes the background for information extraction from Unstructured data i.e data mining ,web content mining .This paper includes available approaches for unstructured data extraction .Then the discussion of the leggings of approaches and concluded with the ideas to overcome the issues.

Keywords-Information extraction; unstructured text;

I. INTRODUCTION

A significant portion of the data on the World Wide Web is in the form of HTML pages. HTML pages are designed to describe the formatting of text, navigational functionality and other visual aspects of a document.[1] About 75% of World Wide Web is of HTML pages. HTML pages are used to carry the format of text and overall visual appearance of the topic. While searching the content of web pages formatting and navigational features should be avoided because they are considered as unwanted data.

Consider contents of book it has table of contents ,editorial images but when we consider search some terms from bunch of eBooks or some digital documents on the internet it becomes a tedious task to separate knowledge from available information. Relational databases have data stored in tables where columns are attributes of table and rows contain data.[5]

Web Mining is the word rooted on data mining. As we know data mining is related to extraction of patterns form data, web mining is related to data on the web. Web mining can be divided in to three categories .Web Usage mining, Web Content Mining, Web url mining. Web content mining is the process of extracting patterns from the unstructured or structured data on the web pages.[4]

II. APPROACHES FOR INFORMATION EXTRACTION

There are several approaches for information extraction from unstructured text.

1. Extracting Unstructured data from template generated web documents.[1]

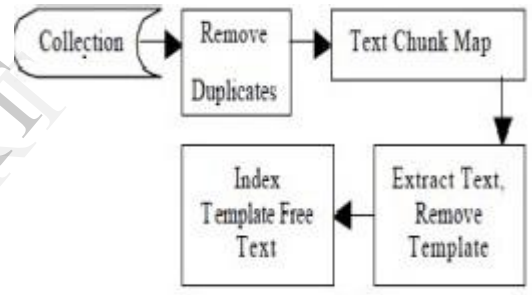

Fig. 1. Text Extraction process

In this approach as shown in figure from a collection of documents first duplicates are removed then text chunk map is created. After duplicates are removed, our algorithm has two passes over the HTML pages.

In the first pass, whenever a delimiter tag or its respective *end tag* such as *table tag* or *image map tag* is encountered, we store the identified text chunk in the *text chunk map* along with the document frequency for that chunk. The document frequency for each *text chunk* is updated while processing the whole collection.

In the second pass, all text chunks that their document frequency is over a determined threshold are identified as *template table text chunk.* The remaining text chunks are the extracted texts that are passed to the *indexer to* be indexed. Thus, the output of the second pass is the extracted text without HTML formatting/ template data.

2. A relational approach to incrementally extracting and querying structure from unstructured data.[2]

Every unstructured document contains a small structure in it. But it cannot be processed directly as any other structured data.

	Jan	Feb	...	Dec
Avg High Temp °F (°C)	23 (-5)	29 (-2)	...	29 (-2)
Avg Low Temp °F (°C)	6 (-14)	12 (-11)	...	13 (-11)
Mean Temp °F (°C)	15 (-9)	20 (-7)	...	21 (-6)
Avg precipitation in (cm)	1.14 (2.9)	1.14 (2.9)	...	1.32 (3.35)

As shown above is a table in a wiki page. It clearly contains tabular structure but we can't query it like a relational table. In this approach with use of wide table, sparse table and complex attributes the data is mapped into a relational like form.

Wide Table

PageId	PageText	History	Economy	temperature
Madison, Wisconsin	"Madison is the captial of the ..."	"Madison was created..."	"Wisconsin state government.."	temperature_wiki

temperature_wiki

City	Month	Low_F	Low_C	High_F	...
Madison, Wisconsin	1	6	-14	23	...
Madison, Wisconsin	2	12	-11	29	...
Madison, Wisconsin	12	13	-11	29	...
Seattle, Washington	1	36	2	46	...

As shown in figure the temperature wiki is a new table created from the will page by avoiding unnecessary symbols. Now we can query this table as relational table and get the information.

3. Automatic segmentation of text into structured records.[3]

In this approach the HMM-Hidden Markov Model[3] is used. A text document containing an address string has structure but there should be some process to extract that structure.

For example ``18100 New Hamshire Ave. Silver Spring, MD 20861" can segmented into five structured elements as follows:

House Number : 18100
Street Name : New Hamshire Ave.
City : Silver Spring
State : MD
Zip : 20861

Steps for above segmentation

1. The _rst step, called Address Elementization , is where addresses are segmented into a fixed set of structured elements.

2. The second step called Address Standardization is where abbreviations (like \Ave.") get converted to a canonical format and spelling mistakes get corrected.

3. Deduplication or Householding phase where all addresses belonging to the same household are brought together. The quality of both these phases can be considerably enhanced by first elementizing the addresses correctly.

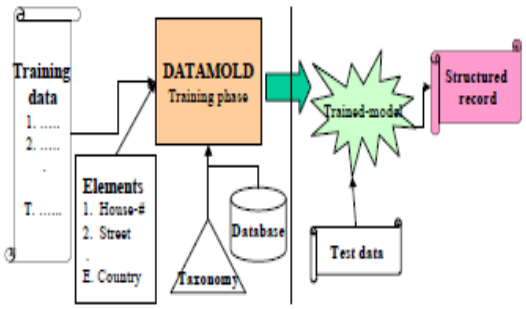

Figure 2: An overview of the working of datamold.

Above figure shows the steps described in the process of this approach

III. CONCLUSION

From all the survey of papers I found a typical generalize approach for information extraction from unstructured text. That is in the below figure.

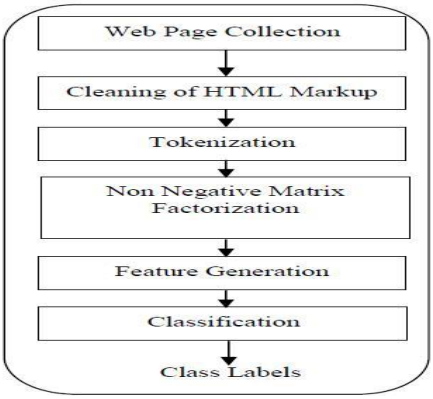

Fig. 3.generalized approach for information extraction form unstructured text[5].

All the approaches for information extraction generally follow the above steps. Feature extraction may have other methods than NMF. From the survey I found an issue for this generalized approach.

i.e. this approach is for web pages. That may changing time to time. so if we want to extract information from the same page after a time of interval and the page have addition of text of a few lines this approache processes the whole page rather than the updated part. So it is waste of time.

There must be a solution for this issue. I have proposed some changes in the above approach for solving this issue.

IV. PROPOSED WORK

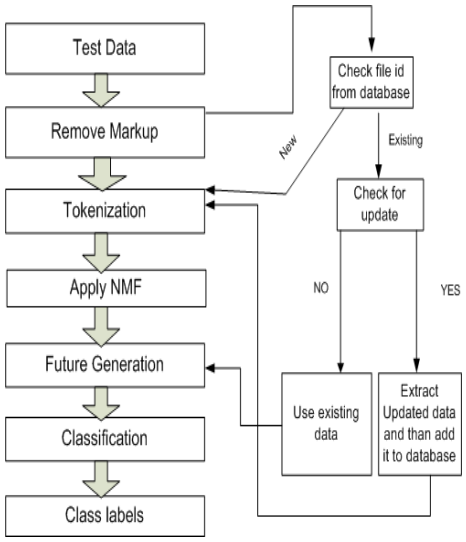

Fig 4. Proposed changes in generalized approach

From above solution the current issues can be solved with some programming implementation. My research paper will be on implementation of this idea. I am currently working on this.

REFERENCES

[1] L. Ma, N. Goharian, A. Chowdhury and M. Chung, "Extracting Unstructured Data from Template Generated Web Documents", *CIKM '03 Proceedings of the twelfth internationalconference on Information and knowledge management*, pp. 512-515,2003.

[2] E. Chu, A. Baid, T. Chen, A. H. Doan and J. Naughton, "A Relational Approach to Incrementally Extracting and Querying Structure in Unstructured Data", *Proceedings of the 33rd international conference on Very large data bases (VLDB '07)*, Vienna, Austria, pp. 1045-1056, September 23-28, 2007.

[3] V. Borkar, K. Deshmukh, and S.Sarawagi ,"Automatic segmentation of text intostructured records", *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, pp. 175 - 186, 2001.

[4] Vidhya. K. and G. Aghila, "Text Mining Process, Techniques and Tools: an Overview", International Journal of Information Technology and Knowledge Management, Vol. 2, No. 2, pp. 613-622, 2010.

[5] Vaishali A. Ingle "Processing of Unstructured data for Information Extraction", 2012 Nirna University International Conference On Engineering NUiCONE-2012.