ABSTRACT

In this paper, we are using boilerplate codes for classifying the individual text elements in a Web page. We enhanced it for date as well as content extraction from a web page by using text analysis process. Similarly classification of web page content is essential to many tasks in web information retrieval such as maintaining web directories, to organize the web information as well as focused crawling. Moreover the algorithm and code has improved for web page categorization. This work is implemented with twenty to thirty web sites and proven that it is correct which can be enhanced more by using web page application. With the help of our model, we proved that web pages categorization and date extraction algorithm is 90 % correct. It categorizes all the web pages and extracts all the relevant dates enclosed within a web page with its relevant content as well as title. Here we are achieving a remarkable accuracy for categorization of web pages, date extraction, title extraction, as well as for comment extraction.

Keywords-- Text mining, Date extraction, Web page categorization. SPO, text analysis

1. INTRODUCTION

When examining a Web page, humans can easily distinguish the main content from navigational text, advertisements, related articles and other text portions. A number of approaches have been introduced to automate this distinction, using a combination of heuristic segmentation and features[1]. However, we are not aware of a systematic analysis of which features are the most salient for extracting those. Similarly classification plays a vital role in many information management and retrieval tasks. On the Web, classification of page content is essential to focused crawling, to the assisted development of web directories, to topic-specific web link analysis, and to analysis of the topical structure of the Web[2]. Web page classification can also help improve the quality of web search. In this paper we analyse the most popular features used for boilerplate detection on two corpora and extended for web page classification. “Web page classification, also known as web page categorization, is the process of assigning a web page to one or more predefined category labels”. Classification is often posed as a supervised learning problem in which a set of labelled data is used to train a classifier which can be applied to label future examples. So here the web page classification and date extraction has been done with the help of open source boiler plate[2] text. Boiler plate is an open source text which is coming in use for the extraction of title as well as relevant content from a web page so it has been extended for extracting the dates by designing an algorithm. The algorithm is designed in regx patterns and implemented with the help of coding. For coding purpose java language and natural language processor are used which is discussed here in this paper. The coding is implemented with twenty to thirty web sites which is 90% correct and it can be extended for more number of websites by enhancing the web page application as well as the date written patterns in the corresponding algorithm. Another concepts which is used here is Spring and Ajax for the development of test applications.

1.1 Text mining

Text mining roughly equivalent to text analytics refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends such as statistical pattern learning. That means deriving patterns within the structured data and evaluation and interpretation of the output. The term text analytics describes a set of linguistic, statistical, and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation.
1.2 Search process optimization

Process optimization is the discipline of adjusting a process so as to optimize some specified set of parameters without violating some constraint. The most common goals are minimizing cost, maximizing throughput, and/or efficiency. This is one of the major quantitative tools in industrial decision making. The goal of search optimization software is to make sure that the search engines deliver a link based on the various parameters available to search along with the content a user wants to look for using Search.

Search process optimization benefits:

- When optimizing a process, the goal is to maximize one or more of the process specifications, while keeping all others within their constraints.
- The purpose of Search process optimization is to allow effective search using Information Extraction not using Information Retrieval on the collection of Unstructured Text.
- Search to use the IE using Text Mining techniques to get the information from the unstructured Text. The Information Extraction steps allow you to configure your search terms based on more parameters available like search using dates, and search between news articles, blogs, forum etc. Information Extraction (IE) systems analyze unrestricted text in order to extract information about pre-specified types of events, entities or relationships.
- Information extraction is all about deriving structured factual information from unstructured text.
- Example, If one want to search where Mr President visited last month, you can’t achieve this, one has to type specifically Mr President and the country name to get the exact result, so one need to aware of the search terms to get the result from the IR.
elements of an HTML page. Other approaches try to exploit structural information, mainly contained in the DOM (Document Object Model) tree of an HTML page, in order to identify common, i.e. frequently used patterns or segments, usually concentrated on specific web sites. (Chakrabarti et. al., 2007) try to determine the “templateness” of DOM nodes, by training Logistic regression classifiers over a set of features, such as closeness to the margins of an HTML page, number of links per word, fraction of text within anchors, size of anchors, etc. At a second level, smoothing is applied on the classifier scores of DOM nodes using regularised isotonic regression to improve performance.

3. RELATED STUDIES

A small set of shallow text features and web page features have been analysed for classifying the individual text elements in a Web page[2]. Moreover, this paper derive a simple and plausible stochastic model for describing the boilerplate creation process and the accuracy graph has been shown by taking Google news website with the help of link density, text density, number of words. The topic of cleaning arbitrary web pages with the goal of extracting a corpus from web data, which is suitable for linguistic and language technology research and development, has attracted significant research interest recently[4]. Several general purpose approaches for removing boilerplate have been presented in that literature. It provides the finer control over the extracted textual segments in order to accurately identify important elements, i.e. individual blog posts, titles, posting dates or comments. BlogBuster tries to provide such additional details along with boilerplate removal, following a rule-based approach. A small set of rules were manually constructed by observing a limited set of blogs from the Blogger and Wordpress hosting platforms. These rules operate on the DOM tree of an HTML page, as constructed by a popular browser, Mozilla Firefox[19] Chekuri et al. studied automatic web page classification in order to increase the precision of web search. A statistical classifier, trained on existing web directories, is applied to new web pages and produces an ordered list of categories in which the web page could be placed. Here the classification category have been studied in the different point of view like binary, multiclass, flat, hierarchical, soft and hard classification.

4. PROBLEM STATEMENT

Boilerplate is an open source text Framework which is in the reference paper using for the title extraction and effective content extraction (the corresponding graph have been shown here) but it does not work for the dates enclosed within a web page. Similarly for web page categorization the algorithms does not work.

![Fig. 3: Text density distribution by class for Google news](image)

In the above graph text Density distribution[2] has been shown with respect to the number of words the distribution function has been shown in form of class for main content, related content and headlines but not for the relevant dates. In this paper we are working for relevant dates as well as work is extended more for categorization.
5. DATE RETRIEVAL

Web page contains the list of items like title, comment, main content and date. Since date is also an important content in a web page but by using boiler plate open source we can extract only the title and main content but not the date so it is extended for date extraction. Dates on a web pages may be present in the different formats so it can be extracted by using date allow filter on boiler plate codes. To extract the date we have used the natural language processor and written it into fixed pattern. With the simulation it has been proved that date extraction algorithm is 90% correct. Documents contain dates explicitly stated in the places like URL, Title, Body of the document, meta tags of the document and Last-modified date from the HTTP response.

The Algorithm extracts all the dates for a document with a matching URL pattern that fits the date format associated with the rule. If a date is written in an ambiguous format, the algorithm assumes that it matches the most common format among URLs that match each rule for each domain that is collected. For this purpose, a domain is one level above the top level. For example, mycompany.com is a domain, but intranet.mycompany.com is not a domain. The algorithm runs a process that calculates which of the supported date formats is the most common for a rule and a domain. After calculating the statistics for each pattern, the process may modify the dates in the data.

Here we work on the Algorithm to extract different dates format present in the body of the content.

5.1 Date retrieval algorithm

In this algorithm the array of months are taken. MM is the numeric month format which is written where square bracket is the optional bracket similarly date patterns DD are written and yyyy/yy is the year format where d stands for digits from 1 to 9. The separator may be dot(,) slash(/) or hyphen(-) which is taken as string.

```
months[]=\{"january","february","march","april","may","june","july","august",
            "september","october","november","december"\};
MM=(0\[1-9]|1[012]);
DD=(0\[1-9]|12|0-9|3\[01]);
YYYY=(1[9]\[1-9]|20\[d\[d]);
YY="(d\[d\[d);"
string [-/]
int numwords;
char/int=month, int MM, int DD, int YYYY/YY var x;
if((x=="today")||(x=="tomorrow")
if(numwords<=4)
{return x;}
else if(x=MM[-/].DD[-/].YYYY)
{Step i) compile this pattern with the pattern object
Step ii) match this pattern with the matcher object
Step iii)return the matched value in solar date format .
Step iv)return x; }
else if(x=MM[-/].DD[-/].YY)
{ go for step i to iv; }
else if(x=DD[-/].MM[-/].YYYY)
{ go for step i to iv; }
else if(x=DD[-/].MM[-/].YY)
{ go for step I to iv; }
else if(x=YYYY[-/].DD[-/].MM)
{ go for step i to iv; }
else if(x=DD[-/].MM[-/].YY)
{ go for step i, ii, iii ,iv; }
else if(x=MM[-/].DD[-/].YYYY)
{ go for step i , ii, iii ,iv; }
```
Classification plays a vital role in many information management and retrieval tasks. On the Web, classification of page content is essential to focused crawling, to the assisted development of web directories, to topic-specific web link analysis, and to analysis of the topical structure of the Web. Web page classification can also help improve the quality of web search. "Web page classification, also known as web page categorization, is the process of assigning a web page to one or more predefined category labels". Classification is often posed as a supervised learning problem in which a set of labeled data is used to train a classifier which can be applied to label future examples. The general problem of web page classification can be divided into multiple sub-problems: subject classification, functional classification, sentiment classification, and other types of classification. Subject classification is concerned about the subject or topic of a web page. For example, judging whether a page is about "arts", "business" or "sports" is an instance of subject classification.

Fig.4: Date Extraction with data cleaning

5.2 Date retrieval pattern samples

Date Format: yyyy-MM-dd
[0-9][4]-[0-9][1-9][1][012][0-9][3][01][0][1-9])
Date Format: yyyy-dd-MM
[0-9][4]-[0-9][1-9][1][012][0-9][3][01][0][1-9])
Date Format: MM-dd-yyyy
1[012][0][1-9][0][1-9][1][012][0-9][3][01][0][1-9][4]"
"MM-dd-yyyy"
Date Format: dd-MM-yyyy
01-9][1][012][0-9][3][01][0][1-9][1][012][0-9][3][01][0][1-9][4]
Date Format: MM-dd-yyyy
[1][0-9][1][012][0-9][3][01][0][1-9][1][012][0-9][3][01][0][1-9][4]
Date Format: dd-MM-yyyy
[1][0-9][1][012][0-9][3][01][0][1-9][1][012][0-9][3][01][0][1-9][4]
classification. Functional classification cares about the role that the web page plays. For example, deciding a page to be a “personal homepage”, “course page” or “admission page” is an instance of functional classification. Sentiment classification focuses on the opinion that is presented in a web page, i.e., the author’s attitude about some particular topic. Other types of classification include genre classification, search engine spam classification and so on.....Based on the number of classes in the problem, classification can be divided into binary classification and multi-class classification, where binary classification categorizes instances into exactly one of two classes multi-class classification deals with more than two classes. Based on the organization of categories, web page classification can also be divided into flat classification and hierarchical classification. In flat classification, categories are considered parallel, i.e., one category does not supersede another. While in hierarchical classification, the categories are organized in a hierarchical tree-like structure, in which each category may have a number of subcategories.

6.1 Page categorization algorithm building blocks

![Page Categorization Algorithm](image)

6.2 Algorithm description

A. Input the system with HTML Text received from the data cleaning
B. Check for the category Matching
   Check for the URL
   a. Match the URL for predefined words for blog, forum, News
   b. If matched Mark the content with the matched word category set
   c. Exit
   d. Check for the Metadata
   e. Identify the content in the Meta data Tag
   f. Match the Meta tag for predefined words for blog, forum, News
   g. If matched Mark the content with the matched word category set
   h. Exit
   i. Check for the Menu
   j. Identify the content in the Menu Tag consisting <li>
   k. Match the Menu tag for predefined words for blog, forum, News
7. SIMULATION OUTPUTS
The software tools JAVA, Spring Framework, GATE NLP and OPEN NLP are used to simulate our work for date retrieval and page categorization.

Date Retrieval:
Simulation is carried on Date Retrieval Algorithm based on the test data for 15 selected random URLs to identify the correctness of Date Retrieval Algorithm.

<table>
<thead>
<tr>
<th>URL</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://javarevisited.blogspot.in/2011/12/difference-between-dom-and-sax-parsers.html">http://javarevisited.blogspot.in/2011/12/difference-between-dom-and-sax-parsers.html</a></td>
<td>Success</td>
</tr>
<tr>
<td><a href="http://javarevisited.blogspot.in/2011/12/difference-between-dom-and-sax-parsers.html">http://javarevisited.blogspot.in/2011/12/difference-between-dom-and-sax-parsers.html</a></td>
<td>Success</td>
</tr>
<tr>
<td><a href="http://javarevisited.blogspot.in/2013/02/disable-submit-button-in-html-javascript-avoid-multiple-form-submission.html">http://javarevisited.blogspot.in/2013/02/disable-submit-button-in-html-javascript-avoid-multiple-form-submission.html</a></td>
<td>Success</td>
</tr>
<tr>
<td><a href="http://timesofindia.indiatimes.com/india/In-12-hours-Mumbai-ladys-credit-card-used-in-4-continents/articleshow/18374084.cms">http://timesofindia.indiatimes.com/india/In-12-hours-Mumbai-ladys-credit-card-used-in-4-continents/articleshow/18374084.cms</a></td>
<td>Success</td>
</tr>
<tr>
<td><a href="http://www.indiatimes.com/">http://www.indiatimes.com/</a></td>
<td>Failure</td>
</tr>
</tbody>
</table>

Table-1

Fig.6: Date retrieval
Page Classification:
Simulation is carried on Page Classification Algorithm based on the test data for 15 selected random URLs to identify the correctness of Page Classification Algorithm.

**Table-2**

<table>
<thead>
<tr>
<th>URLs</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://javarevisited.blogspot.in/2011/12/difference-between-dom-and-sax-parsers.html">http://javarevisited.blogspot.in/2011/12/difference-between-dom-and-sax-parsers.html</a></td>
<td>Success</td>
</tr>
<tr>
<td><a href="http://javarevisited.blogspot.in/2011/12/difference-between-dom-and-sax-parsers.html">http://javarevisited.blogspot.in/2011/12/difference-between-dom-and-sax-parsers.html</a></td>
<td>Success</td>
</tr>
<tr>
<td><a href="http://javarevisited.blogspot.in/2013/02/disable-submit-button-in-html-javascript-avoid-multiple-form-submission.html">http://javarevisited.blogspot.in/2013/02/disable-submit-button-in-html-javascript-avoid-multiple-form-submission.html</a></td>
<td>Success</td>
</tr>
<tr>
<td><a href="http://timesofindia.indiatimes.com/india/In-12-hours-Mumbai-ladys-credit-card-used-in-4-continents/articleshow/18374084.cms">http://timesofindia.indiatimes.com/india/In-12-hours-Mumbai-ladys-credit-card-used-in-4-continents/articleshow/18374084.cms</a></td>
<td>Success</td>
</tr>
<tr>
<td><a href="http://www.indiatimes.com/">http://www.indiatimes.com/</a></td>
<td>Failure</td>
</tr>
</tbody>
</table>
8. CONCLUSION

Text mining is a burgeoning technology, because of its newness and intrinsic difficulty, in a fluid state, perhaps, to the state of machine learning. When the term is broadly interpreted, many different problems and techniques come under its ambit. In most cases it is difficult to provide general and meaningful evaluations because the task is highly sensitive to the particular text under consideration. Document classification, entity extraction like date, and filling templates that correspond to given relationships between entities, are all central text mining operations that have been extensively studied. Using structured data such as Web pages rather than plain text as the input opens up new possibilities for extracting information from individual pages and large networks of pages. Automatic text mining techniques have a long way to go before they rival the ability of people, even without any special domain knowledge, to glean information from large document collections. This work extended further to extract the relevant information and filtering out the comment section with the sentiments i.e whether the comment is positive or negative. The performance graph can be extended by defining codes for number of date patterns i.e design the code for all the predefined date formats. The future work may also be extended for the specific content search. Webpage classification can help improve the quality of web search and improve SEO skill.

References


and Evaluation (LREC’08), Marrakech, Morocco, may 2008. European Language Resources Association (ELRA).


AUTHOR

Lambodar Jena received his M.Tech degree in Computer Science and Application from Indian School of Mines(ISM), Dhanbad in the year 2000 and currently pursuing doctorate(Ph.D) degree from Utkal University, India. His research interest includes data mining, data privacy, image mining, web mining and wireless sensor networking. He has several publications in national/international journals and conference proceedings.

Narendra Kumar Kamila received his M.Tech degree in Computer Science and Engineering from Indian Institute of Technology, Kharagpur and doctorate(Ph.D) degree from Utkal University, India in the year 2000. Later he had visited USA for his post doctoral work at University of Arkansas in 2005. His research interest includes artificial intelligence, data privacy, image processing and wireless sensor networking. He has several publications in journal and conference proceedings. His professional activities include teaching computer science, besides he organizes many conferences, workshops and faculty development programs funded by All India Council for Technical Education, Govt. of India. Dr. Kamila has been rendering his best services as editorial board member and editor-in-chief of many international journal and conference proceedings.