ISSN: 2349-7300
ISO 9001:2008 Certified
International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences (IJIRMPS)
Volume 2, Issue 2, April 2014

Web Content Extraction Using Machine Learning

Prof. Yagnesh Tiwari

Assistant Professor
L.J. Institute of Technology, Ahmedabad

***Abstract*: Extraction model aims at separating the main data from noise. We define content as a continuous and meaningful resource of text from web pages which can be successfully used to summarize required topic in a concise way. Noise can be any parameter of a web page. Noise on the other hand is defined by any web page parameter that deviates from the main content. Noise can be copyright disclaimers, advertisements, navigation arc. Boilerplate templates form a major extraction criteria of Bolierplate detection algorithm.**

Descriptors of the domain: Information Systems, search, content retrieval.

***Keywords*: Content Extraction, Boilerplate Removal, Template Detection, CETR.**

I. Introduction:

Information systems have seen a great boom in the computing industry lately. With internet being the largest information resource, extraction of information from various sources on the internet has become an increasingly challenging and most demanded process by most systems. Extensive research has been done on the extraction models. These models are used by many automated systems and crawlers for effective extraction of information adequately fitting to requirements at hand.

Web pages are quite dynamic in nature consisting of many graphical information amongst which is embedded continuous lengths of text. This forms the prime key to our extraction model. A web site or webpage at hand can be broken down to blocks. These blocks constitute of similar items as navigation elements, advertisements and main content itself.

The goal of our thesis is to create a language independent extraction model to separate this main content from any deviation of topic in search which we define as noise using supervised decision trees [1] and unsupervised K-means [4] machine learning algorithms.

II. Problem Statement:

The problem of content extraction has been worked on extensively but most extraction models deal with the extraction of structured data. Our thesis aims at creating a model that extracts and summarizes content from unstructured data. Unstructured data can be defined as continuous blocks of texts on the web pages we aim to derive meaningful information from ad thus summarize as per the requirements at hand.

III. Working

1. **Input:** Web page url.
2. Crawler recursively parses links in the web page source html.
3. Dom parsed by the crawler is passed on to extraction model.
4. Content Extraction algorithms:

a) Boilerplate Detection: Blocks are identified using screen layout altering tags such as <h1>,<p>, .Blocks detected are tokenized and compared with corresponding gold standard dataset for labeling and further extraction through decision trees.

b) CETR: Content extraction through tag ratios makes use of unsupervised K-Means algorithm for extraction.

ISSN: 2349-7300
ISO 9001:2008 Certified
International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences (IJIRMPS)
Volume 2, Issue 2, April 2014

(Semantic tags have been introduced as additional feature sets for above method).

5. **Output:** Output is received in the form of summarized and noise separated continuous text of relevant information in question.

The algorithm uses two major approaches for extraction of content .The very first approach is Boilerplate detection [1].The second approach is the extractor through tag ratios [4].Tag ratios is a subsidiary methodology in our extraction model, the reason for the same has been explained further.

Boilerplate is a template of any webpage. Boilerplate detection the key to content extraction; this can be simply understood as templates of any webpage hold structural key to the content.

Content is embedded in the webpage's boilerplate template. Various templates are provided by multiple web design platforms online which comprises of the major portion of web content. This further supports our cause by maximizing the reach of scope of our extractor model.

Also our approach being based on machine learning a strong and uniform history of webpages increases accuracy of our model. Boilerplates that have already come across on the training model will produce faster and more accurate results against content that is comparatively newer and have not been tested by our model.

Second approach is based on the tag ratios. Tag ratios are of html tags to continuous text characters in the source html lines. Both the mentioned approaches primarily aim at dividing the input source html into blocks .Boilerplate detection divides the pages into blocks on the basis of structural altering tags such as <h1>,<p>, etc. The second approach on the other hand uses line breaks as division for the blocks.

The second method used html tag to content text ratio as a measure for classification. This proves to be an ineffective approach for newer datasets and modern webpage designs which are dynamic in nature. The styling CSS elements of the webpage are straight away discard in this approach which are key elements of modern web pages.

To resolve the above conflict and still preserve the essence of this technique we introduce tag ratios as a feature set to the boilerplate detection thus resulting in better extraction results. We further introduce semantic html features which are key directives in identifying content and separating from noise.

IV. Dataset:

Our primary data set was collected in late 2012 and consists of English language pages from three sources: i) 999 pages randomly selected from popular RSS feeds with a large number of subscribers; ii) 204 pages from a selection of 23 large news sites (e.g. bbc.com and nytimes.com);and iii) 178 pages randomly sampled from a blog directory (technoratti.com) across all categories and authorityranges.3The gold standard was extracted with a web browser by pasting it into a text file. We considered any article text including title, date and author information as well as any user generated comments to be content. To benchmark against previous studies, we also use a portion of the data set from [4]. This includes both the Cleaneval data, as well as data from [2].

V. Machine Learning Approach:

Our approach primarily aims at splitting the web page at hand into blocks. We have combined two machine learning approaches for effective extraction of data. Both the methodologies aims at splitting the given data into blocks

Boilerplate Detection [1] aims at splitting the DOM [3] into blocks and using supervised decision trees. CETR [4] aims at splitting the data to be extracted as per line breaks.

Both the approaches assume the content blocks to be more dense than the non-content blocks, however the detailed working of the same vary greatly.

The boilerplate removal technique [1] begins with parsing the DOM [3] for splitting same into blocks. This is done iteratively looking for tags like ,<p>,<h1>.These tags and other similar tags are chosen because they are suspected for on screen layout modification. Each of these tags are further split into individual blocks. Blocks with no content at all are discarded.

Now the algorithm need to label these blocks. For labelling we assign tokens to the content the blocks. Examples of tokens are **nav, menu, widget, title, header, facebook, twitter, comment, author, thread etc.** Each token in the entire document is associated to a single block

Next we apply longest common subsequence to test the percentage of tokens of each block against that of percentage in the gold standard dataset.

Any block with more than 10% of the tokens extracted is tagged "content". The data is randomly sampled into 70%/30% training/test split. We use L2 regularized logistic regression, with the regularization parameter set via 5-fold cross validation in the training data.

ISSN: 2349-7300
ISO 9001:2008 Certified
International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences (IJIRMPS)
Volume 2, Issue 2, April 2014

Table 1 gives a comparison of sample tokens tested against gold standard and their results in percentage of blocks.

Table 1.Comparison of tokens in class attributes:

TOKEN	CONTENT:NO CONTENT	PERCENT OF BLOCKS
Menu	1 : 373.6	2.2%
Widget	1 : 314.1	4.6%
Nav	1 : 68.9	3.3%
facebook	1 : 18.3	1.3%
Top	1 : 13.3	1.9%
twitter	1 : 8.5	2.3%
Title	1 : 3.3	10.5%

V. Results

Combination of features has resulted in a favourable improvement of results. Semantic features introduced give us leverage over the newer datasets in performance. Splitting the feature sets has not resulted in an equally good performance as combined. This probably is because newer html sources have strong dominance of CSS features. Semantic tags have been seen to efficiently use the information left behind by programmers to extract data for the combined models making it a better option for newer datasets. Also previous features alone extract the content from tokens in blocks .Blocks have been created identifying the screen altering tags such as <p>, ,<h1> etc. However practical observation of real time web page sources reveal that a lot of relevant content may be hidden in higher levels of DOM, thus decreasing the performance on the boilerplate detection technique alone.

INPUT:

OUTPUT:

ISSN: 2349-7300
ISO 9001:2008 Certified
International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences (IJIRMPS)
Volume 2, Issue 2, April 2014

Table 2 gives a overview of the accuracy in terms of mean F1 scores for various feature set combinations that we have used in our model.

Table 2.Model comparison of mean F1-Score:

Feature Combinations	Cleaneval-EN	BIG 5	2012 Train	2012 TEST
Baseline	0.899	0.625	0.621	0.623
CETR	0.919	0.794	0.741	0.725
IC	0.887	0.641	0.709	0.701
ST	0.904	0.854	0.817	0.809
ST+IC	0.896	0.858	0.836	0.824
ST+IC+CETR	0.907	0.887	0.848	0.836

V. Conclusion

The results obtained through these models are primarily basic but introduction of semantic features and combining new features improves the performance of extraction model. Benchmark performances are obtained for combination of various features. Future scope for the system has high potential for applications in data archiving and information systems with increased boom of data analytics and related applications.

References

[1] C. Kohlsch¨utter, P. Fankhauser, and W. Nejdl Boilerplate detection using shallow text features. In Proceedings of WSDM ’10, pages 441–450. ACM, 2010.
[2] J. Pasternack and D. Roth. Extracting article text from the web with maximum subsequence segmentation. In Proceedings of WWW ’09, pages 971–980. ACM, 2009.
[3] F. Sun, D. Song, and L. Liao. Dom based content extraction via text density. In SIGIR, volume 11, pages245–254, 2011.
[4] T. Weninger, W. H. Hsu, and J. Han. CETR: content extraction via tag ratios. In Proceedings of WWW ’10,pages 971–980 .ACM,2010