Web Content Extraction Using Hybrid Approach

Dhumal Tania, Kumbhar Shital, Malave Sumedha, Salunkhe Shrutika

Computer Engineering, Trinity Academy of Engineering, Pune, Maharashtra, India

ABSTRACT

Wide Web has rich source of voluminous and heterogeneous information which The World continues to expand in size and complexity. Many Web pages are unstructured and semi-structured, so it consists of noisy information like advertisement, links, headers, footers etc. This noisy information makes extraction of Web content tedious. Extracting main content from web page is the preprocessing of web information system. Many techniques that were proposed for Web content extraction are based on automatic extraction and hand crafted rule generation. A hybrid approach is proposed to extract main content from Web pages. A HTML Web page is converted to DOM tree and features are extracted and with the extracted features, rules are generated. Decision tree classification and Naive Bayes classification are machine learning methods used for rules generation.

Keywords: Website, Automatic Extraction, Handcrafted Rules, Hybrid Approach, Machine Learning.

I. INTRODUCTION

WWW allow to upload and download relevant data and valuable contents through websites. Data is in unstructured or semi-structure so lot of irrelevant document obtain after navigating several link. So data mining technologies can not apply directly. For effective retrieval of web information called “Web Mining”. After introduction of web mining we use data mining technology. After certain stage using clustering and classification extraction of original contents are is not possible so we use technology like handcraft, DOM(Document Object Model) classification which are group in to two sections-1) Automatic 2) handcrafted. In this we use FE(Feature Extraction) also and in feature extraction TD (Table Data) and div tag also use. With the help of machine learning we can increase the performance of machine. It also consist of Decision tree classification and Naive Bays classification. After performing this all operation unwanted advertise will be remove successfully and we all get only plain text.

II. RELATED WORK

Existing Web Content Extraction techniques are grouped into two major categories (i) Automatic Extraction, (ii) Handcrafted rules generation.

2.1 Automatic Extraction

Automatic Extraction is the process of extracting the Web page content automatically using tools and techniques. Web page segmentation can be done based on three approaches and they are DOM-based segmentation, location-based segmentation and visual-based segmentation.

2.2 Handcrafted Rules

Hand crafted rule generation uses string manipulation function for rule generation. Hand-crafted rules are impractical for more than a couple of data source.

III. HYBRID APPROACH

![Diagram of Hybrid Approach](image-url)
3.1.1 DOM tree Construction

To demonstrate the visual content richer features of a Web page, a hierarchy called DOM is used.

![DOM Tree Construction](image)

**Figure 2. DOM Tree Construction**

IV. FEATURE EXTRACTION

4.1 Best First Method

It searches the space of features subset by gridy hill climbing with backtracking facility.

4.2 Greedy Stepwise Method

It is used to perform FCFB search and it is based on who come first.

4.3 Rankers Method

In this Method individual programming will done.

V. APPLYING MACHINE LEARNING METHODS

Machine learning is a process by which a system improves its performance. Two Machine learning technique’s like decision tree classification and Naïve Bayes classification is used to extract rules.

5.1 Decision-Tree Classification

![Decision-Tree Classification](image)

5.2 Naïve Bayes classification

![Naïve Bayes classification](image)

VI. ARCHITECTURE VIEW

![Architecture View](image)

VII. LITERATURE SURVEY

7.1 Existing System

7.1.1 Template Detection (TD)
- Time consuming process to extract content from web mining.

7.1.2 Machine Learning
Delay report.
Creating multiple identities.

7.1.3 Fuzzy Association rules (FAR)
- Tag identification is not easy.
- Expected output is not come even if it will take longer time for identification.
- Because of this drawbacks the developers, mathematicians and researchers
  have to face many problems in their task.

7.2 Proposed System

7.2.1 Template Detection (TD)
- They demonstrated that a well-chosen combination of different content extraction algorithms can provide better results than a single approach on its own.

7.2.2 Machine Learning
- Composite text density which include
  1. LinkCharNumber
  2. LinkTagNumber

7.2.2 Fuzzy Association rules (FAR)
- Kohlschutter et al. developed a simple, yet effective technique to classify individual text elements from a web page.

8. System Design

VIII. MATHEMATICAL MODEL

The source of the page is denoted as

The source page is split in to various segments as shown in equation (1)

\[ \Omega = \{ !1!2!3!4, \ldots, n \} \cdot \cdot \cdot \cdot \cdot \cdot \cdot (1) \]

In (1) each represents a segment of the web page.

The text contents of are separated from the html tags as shown in equation (2)

\[ = 8 = \{ ini = 1.. : (!) \} \cdot \cdot \cdot \cdot \cdot \cdot \cdot (2) \]

Represents the function to strip textual contents from the html tags. This step is performed to make the content analyzer to consider only the textual contents and omit the tags which are used for formatting the contents.

After the removal of tags, the contents are submitted to content analysis service. The content analysis service returns an array which holds both the significant terms and their weight, as shown in equation (3)

\[ = \{(W_i)\} \cdot \cdot \cdot \cdot \cdot \cdot \cdot (3) \]

represents the Yahoo!

The user’s profile bag is represented with a set of keywords as shown in equation (4)

\[ \Gamma = \{ b_1, b_2, \ldots, b_n \} \cdot \cdot \cdot \cdot \cdot \cdot \cdot (4) \]

The segments of the page weighed against these profile keywords b1.

MATHEMATICAL MODEL

The various dimensions with which the segments are evaluated with the profile keywords \{L, I, V, T\} - L indicates Link, I indicates Image, V indicates Visual Weight and T indicates Theme weight

IX. RESULTS AND DISCUSSION

Machine learning methods like Naïve Bayes classification and C4.5 decision tree classification. From that rules are generated and using the rules informative content of the Web page is extracted. Performance of Naïve Bayes classification and C4.5 decision tree classification method is obtained by
calculating the metrics. Metrics like precision, recall, F-
measure and accuracy are

![Performance Evaluation based on Precision](image)

**Figure 1.** Performance comparison based on precision

The Fig., gives the comparison of C4.5 decision tree
classification and Naïve Bayes classification based on
the metric precision. When numbers of words in a
HTML file increases, precision of C4.5 decision tree
classification is high when compared with Naïve Bayes
classification. C4.5 decision tree classification achieves
89% precision whereas Naïve Bayes classification
achieves only 81% precision.

**X. CONCLUSION**

We can develop an application which can remove spam
adverts. A Web page is converted to DOM tree and
features are extracted.

**XI. REFERENCES**

[1]. S. Baluja, Browsing on smalls screens: Recasting
Webpage segmentation in to an efficient machine
learning framework, Proceedings of the 15th
3342, 2006.

[2]. S. Debnath, P. Mitra, N. Pal and C. L. Giles,
Automatic identification of informative sections of
Web pages, IEEE Transactions on Knowledge and
Data Engineering, Vol. 17, No. 9, pp. 12331246,
2005.

[3]. S. Mahesha, M. S. Shashidhara and M. Giri, An
Efficient web content extraction using mining
techniques, International Journal of Computer
Science and Management Research, Vol. 1, No. 4,

[4]. Nikolaos Pappas, Georgios Katsimpras and