Appling Link Target Identification and Content Extraction to improve Web News Summarization

Rodolfo Ferreira
Rafael Ferreira
Rafael Dueire Lins
Hilário Oliveira
UFPE, Recife, PE, Brazil
rodolfoferreira@gmail.com

Marcelo Riss
HP Brazil R&D
Porto Alegre, RS, Brazil
marcelo.riss@hp.com

Steven J. Simske
HP Labs.
Fort Collins, CO 80528, USA
steven.simske@hp.com

ABSTRACT

The existing automatic text summarization systems whenever applied to web-pages of news articles show poor performance as the text is encapsulated within a HTML page. This paper takes advantage of the link identification and content extraction techniques. The results show the validity of such a strategy.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text analysis.

Keywords

Summarization; Content Extraction; Link Identification

1. INTRODUCTION

The Internet or the World Wide Web is an important source of information today. A web page is complex HTML structure in which the text is mixed with codes, images, links to other web pages, etc. Several systems for web page summarization may be found in the literature, as example [1]. Although, those systems target the extractive summarization of the text part of web pages, they have two drawbacks: (i) They do not separate web pages with relevant information from the ones with advertising and directories; (ii) They are too general purpose, as they are not designed to a specific type of text, such as news articles.

This paper presents a new approach to automatically perform the extractive summarization of web pages of news articles. The proposed system receives as input the root link of any news site and it identifies and summarizes all news. This process is performed in four steps: Web Crawler, Link Target Identification, Content Extraction, and Text Summarization.

The experimental results obtained for web pages show that the proposed system reaches comparable results to the summarization of the plain text (with the text extracted manually) of the same web pages. This proves the efficiency of the new summarization system proposed. The link identification/ classification step achieved 96.72% and 96.70% in terms of accuracy and F-measure, respectively; and the proposed content extraction algorithm outperforms the state-of-the-art system by 8.35% using F-measure.

2. BACKGROUND

This section presents an overview of the main concepts associated with the proposed system.

2.1 Link Target Identification

The link analysis field studies the relationship between web pages. This research area flourished after the proposal of the page rank algorithm [12], which assesses the importance of a web page according to the number of external references (links) to it. Some other works follow the same direction [16]; none of them use the link as a way to classify the content of web pages, however.

The link analysis proposed here, called link target identification, is a completely original step. It benefits from several features extracted from the link structure in order to classify the link-target page as a “news article" or an “irrelevant page" (advertisement, non-text information, etc.).

2.2 Content Extraction Methods

The Content Extraction (CE) process tries to detect and extract relevant textual content from web pages, avoiding navigation menus, advertisement banners, and structural information.

Document Slope Curves (DSC) [13] treats a web page as a sequence of tokens, which could be HTML tags or text, looking for sections with a low number of HTML tags.

The Texto To Ratio (TTR) algorithm [14] counts the number of HTML-tags per line. Lines with a high number of text characters are considered relevant content.

Link Quota Filter (LQF) could be used as a CE algorithm and as a preprocessing step for other CE algorithms. The algorithm looks for the ratio of text which is inside the DOM-tree nodes to text outside of these nodes [4].

The Boilerplate Detection algorithm [7] uses text features, structural information and densitometry features to find the content sections of a website. Decision trees and linear support vector machines are used to classify the sections into content or boilerplate.

The drawbacks of the recently proposed content extraction algorithms [6] are: (i) They are supervised, requiring an annotated corpus to perform the extraction; (ii) They were developed to improve the printing process; thus, the images within the web page can also be selected as relevant content; (iii) They are slower than the solution proposed here.

2.3 Extractive Summarization Methods

An extractive summarization platform selects from the original text document a subset formed with the sentences considered as

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

DocEng'16, September 13–16, 2016, Vienna, Austria.
© 2016 ACM. ISBN 978-1-4503-4438-8/16/09...$15.00
DOI: http://dx.doi.org/10.1145/2960811.2967158

most relevant exactly as they appear, to form the summary. Ferreira et al. [2] evaluated the 17 extractive summarization methods described in the literature, of which reference [3] points out the six sentence scoring methods better suitable for news articles:

- **Word Frequency**: The more frequently a word appears in a text, the higher would be its score.
- **TF/IDF (Term Frequency/Inverse Document Frequency)**: Formula 1 gives scores to sentences.

$$TF/IDF(w) = (NumTS) * log(\frac{(NumS)}{(MenST)}) \qquad (1)$$

  NumTS = frequency of term t in sentence S, NumS=total number of sentences, MenST = sentences with term t.

- **Lexical Similarity**: Relates sentences that employ words with the same meaning (synonyms) or other semantic relation.
- **Sentence Position**: The most important sentences tend to appear at the beginning of a document.
- **Sentence Resemblance to the Title**: the vocabulary overlap between a sentence and the document title increases its importance.
- **Sentence Length**: short and long sentences are penalized.

3. WEB NEWS SUMMARIZATION

The input of the system is a root link to some website of news articles, such as CNN[1]. The target web page is processed in four steps before reaching the final output that is the summarization of all news found in the website. The following sections describes each step.

3.1 Link Target Identification Step

The goal of the link target identification step is to classify links as “news article" or “irrelevant page". Thus, it eliminates pages containing advertising, navigation menu, videos, etc.

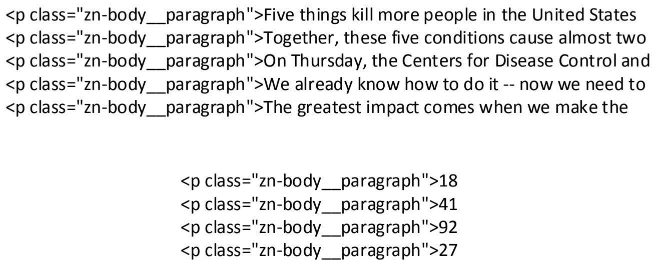
The proposal is to create a classification to retrieve the relevant content pages, more specifically a news article web page. Six features based on link structure are used:

- **Has Number**: This attribute checks if the link has some number in the “name" of the link. Links related to news usually have an ID in its structure;
- **Has Date**: If the link contains a date, it is more likely it to be a news article. In general, advertising and navigation menus have no date in they link structure;
- **Length**: Usually longer link structures are news. Some websites use the title of the article in the link;
- **Ends with Slashes**: The attributes responsible to divide a navigation menu usually end with a slash sign in the case of a news article web page;
- **Has reserved Word**: This attribute was chosen to filter multimedia news, as the proposed system targets text. The proposed system adopts the words *gallery, video, image, photo, slideshow, episode* and *player* as a list of reserved words indicating that the web page pointed at by such link cannot be processed;
- **Slash number**: Count the number of slashes in the link address, indicating the depth of the page in the website.

Several different supervised machine learning algorithms were tested to automatically classify the proposed features of links. Further details on the classification and the training corpus are presented in Section 4.1.

3.2 Content Extraction Step

Once the link is classified as pointing at a text news article web page, the next step is to extract the relevant content from the HTML page provided by the link. Initially, the proposed algorithm processes the web page to remove the native HTML attributes that are not used for text content, some examples are *style, color, padding, margin* and *likewise*. Then the content in each tag is replaced by its number of words. In other words, it creates a new document only with the tags and the number of words inside each tag, as shown in Figure 1. After the initial processing the proposed algorithm creates a DOMTree based on the remaining tags. It contains the number of words of each tag, as presented in Figure 2. Once the DOMTree is completely built, the system scans every leaf in order to remove void tags or script codes. Those tags are used to: (i) detail the style of some text, (ii) group other tags inside it, and (iii) encapsulate code (program) inside the web page.



```
<p class="zn-body__paragraph">Five things kill more people in the United States
<p class="zn-body__paragraph">Together, these five conditions cause almost two
<p class="zn-body__paragraph">On Thursday, the Centers for Disease Control and
<p class="zn-body__paragraph">We already know how to do it -- now we need to
<p class="zn-body__paragraph">The greatest impact comes when we make the
```

```
<p class="zn-body__paragraph">18
<p class="zn-body__paragraph">41
<p class="zn-body__paragraph">92
<p class="zn-body__paragraph">27
```

Figure 1: Content Tag with Text and Numbers

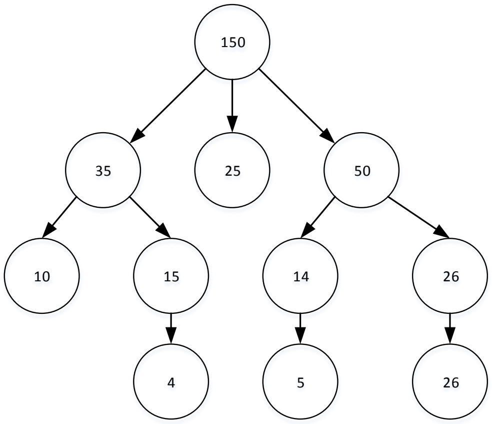


Figure 2: Initial Tree

Then, the algorithm checks the number of words inside the nodes, if it has at least ten words, then it is a possible candidate

[1] www.cnn.com

to point at a news relevant content; otherwise, the system excludes the node, as presented in Figure 3. The last step is the core of the content extraction step. It follows a bottom-up approach that calculates the sum of the number of words in children nodes and in their parent node. The sytem tests if the number of words in the parent node is less or equal to the sum of the words in its children (difference <=0). If so, its child nodes are deleted; otherwise they are kept. Figure 4 details this step.

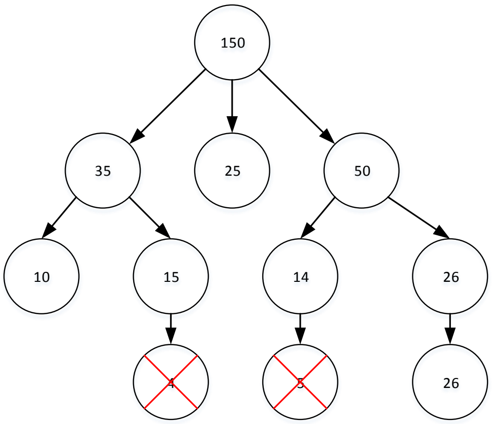


Figure 3: Eliminating nodes with less than ten words

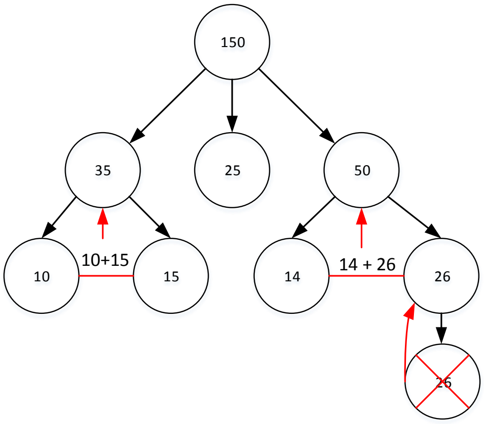


Figure 4: Final Tree

3.3 The Text Summarization Step

The proposed system creates two ordered vectors to represent the text (first step of text summarization). The vectors contain the list of sentences and words from the text that will be the input to the sentence scoring process.

The system employs six sentence scoring services that meet the specifications detailed in Section 2.3. It is important to observe that all those services provide an output score between 0 and 1 for each sentence.

The scores obtained by the summarization methods are averaged to find the most relevent sentences of each text. The sentences with the highest scores compose the summary.

4. EVALUATION

The evaluation of the proposed platform was divided in three parts as follows.

4.1 Link Target Identification

A dataset containing 1,590 instances was created. The links were collected from 10 different news websites (ABC news, BBC news, CBC news, CNN, Daily post, Fox news, Yahoo news, the New York times, the Verge, Reuters), of which 800 were tagged as news. Different machine learning algorithms were applied in order to find which one better fits the proposed PI approach. The WEKA Data Mining Software [15] was used.

Table 1 presents the results of the proposed method using the classifiers listed. The data obtained show that the worst classifier reached 93.45% of Accuracy and 93.50% of F-measure. Thus, these results confirm that the proposed set of features achieve good results in the task of identifying news article web pages by inspecting only the “name of the link".

Table 1: *Results of the proposed approach*

Algorithm	Accuracy	F-measure
Bayesian Network	93.45	93.50
MLP	96.10	96.10
C4.5	96.72	96.70
KNN	96.41	96.40

Among those algorithms, the C4.5 classifier was chosen for the following reasons: (i) It achieves better results; (ii) The output of C4.5 is human interpretable; (ii) It generates a tree that can be easily translated into a set of rules.

4.2 Content Extraction

A dataset extracted from the CNN website was used to perform the evaluation on the Content Extraction and Summarization steps.

The CNN-corpus [9] encompasses 3,000 texts in ten different subject categories, originally tagged by CNN. The CNN-corpus dataset was created to assess automatic summarization methods; however, since its source is news articles from the CNN website it was easily adapted to the content extraction task.

Gottron [5] presents four different levels of granularity of representations to evaluate CE algorithms. This work adopted the words as a set approach. Such a method provides results very similar to its competitors, but is simpler to implement and evaluate the results obtained.

Table 2 presents the results for each related work against the new approach in terms of precision, recall and F-measure. The proposed system reaches a result 17.40% and 8.35% better than its competitors in terms of recall and F-measure, respectively. In relation to the precision, the DSC + LQF algorithm achieves the best results.

Table 2: *Evaluation of Content Extraction Methods*

Algorithm	Precision	Recall	F-measure
Proposed System	50.13	**71.37**	**58.90**
Boilerplate	49.17	60.79	54.36
DSC + LQF	**58.57**	41.63	48.66
TextToRatio	08.54	05.18	06.44

The main interpretation of these results is that the proposed system retrieves 17.40% more relevant content than the others (recall), even with a lower precision than DSC + LQF. As the main goal of the system is text summarization, it is more desirable to have more text with a little noise, than losing some important parts of the text.

4.3 Text Summarization

The main goal of this evaluation is to assess the summarization results using text-only files and a complete HTML page. Thus, the results compare: (i) the text summarization system presented by [3] and (ii) the proposed system that performs the content extraction phase to identify the main content.

The evaluation of the quality of the final summaries was performed using ROUGE [8] on the CNN dataset. ROUGE-1 measure was adopted in order to show the relation among the summaries created from a simple text data and the text extracted from web pages (proposed system). It is important to notice that both systems use the same summarization method, the difference here being only the input file (simple text and HTML page).

Table 3 shows the results using ROUGE 1, in which one may see that the proposed system achieves a better precision when compared with the original web page, thus applying the process described in this paper yields summaries that contain more relevant words than the traditional summarization methods [2]. On the other hand, it achieves a lower recall which shows that some relevant words were not recovered by the system. The results of the F-measure show that the traditional system reaches less than 1% improvement, being statistically equivalent. Some conclusions may be drawn: (i) The ROUGE 1 results show that the proposed system retrieves as many important concepts as the traditional summarization system. This measure assesses the number of unigrams in the generated summary that is in the gold standard summary; (ii) The results of ROUGE 1 show the equivalence between the system proposed and the traditional system.

Table 3: *Summarization Evaluation - ROUGE 1*

System	**Precision**	**Recall**	**F-measure**
Proposed System	47.49	55.18	51.05
Text-only Web page	46.62	58.11	51.73

5. CONCLUSIONS AND FURTHER WORK

This paper proposes a system to summarize news article web pages, which takes advantage of a new link target identification classifier and a new content extraction algorithm proposed here.

The link target identification classifier uses six features based on link structure to classify the HTML page as “news article" or “irrelevant page". The best classifier proposed reached 93.45% and 93.50% of Accuracy and F-measure respectively.

The content extraction algorithm proposed uses the HTML DOM tree to identify the main content of the HTML page. The evaluation in the proposed content extraction algorithm shows a result 17.40% and 8.35% better than the competitors in terms of recall and F-measure, respectively.

The assessment the summarization process showed that the proposed system retrieved as many important concepts as the same summarization methods applied to text-only files.

Currently, these following research lines are under development: (i) Evaluating the proposed system using different datasets; (ii) Creating a web crawler based on the news summaries; (iii) Extending the approach presented here to other kinds of web pages.

Acknowledgments

The research results reported in this paper have been partly funded by a R&D project between Hewlett-Packard do Brazil and UFPE originated from tax exemption (IPI - Law number 8.248, of 1991 and later updates).

6. REFERENCES

[1] N. Akhtar, B. Siddique, and R. Afroz, “Visual and textual summarization of webpages,” in *Data Mining and Intelligent Computing*. IEEE, 2014, pp. 1–5.

[2] R. Ferreira *et al*. Assessing sentence scoring techniques for extractive text summarization. *Expert Systems with Applications*, 40(14):5755–5764, 2013.

[3] R. Ferreira, *et al*. A context based text summarization system. In *Document Analysis Systems*, pp. 66–70. IEEE, 2014.

[4] T. Gottron. Evaluating content extraction on HTML documents. In *International Conference on Internet Technologies and Applications*, pp. 123–132, 2007.

[5] T. Gottron. Evaluating content extraction on HTML documents. *ITA*, pp. 123 – 132, 2007.

[6] T. Hassan and N. Damera Venkata. The browser as a document composition engine. In *Symposium on Document Engineering*, pp.pp 3–12. ACM, 2015.

[7] C. Kohlschutter, P. Fankhauser, and W. Nejdl. Boilerplate detection using shallow text features. In *Web Search and Data Mining*, pp. 441–450, 2010.

[8] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In ACL-04 Workshop, pp. 74–81,2004.

[9] R. D. Lins, *et al*. A multi-tool scheme for summarizing textual documents. In *WWW/INTERNET 2012*, pp. 1–8, 2012.

[10] E. Lloret, M. T. RomÃ¡-Ferri, and M. Palomar. Compendium: A text summarization system for generating abstracts of research papers. *Data & Knowledge Engineering*, 88(0):164-175, 2013.

[11] A. Nenkova and K. McKeown. A survey of text summarization techniques. In *Mining Text Data*, pp. 43–76, 2012.

[12] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[13] D. Pinto *et al*. Quasm: A system for question answering using semi-structured data. In *Joint Conference on Digital Libraries*, pp. 46–55, 2002.

[14] T. Weninger, W. H. Hsu, and J. Han. Cetr: Content extraction via tag ratios. In *World Wide Web*, pp. 971–980, 2010.

[15] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Pub. Inc., 2000.

[16] Z. Zhao, *et al*. Topic oriented community detection through social objects and link analysis in social networks. *Knowledge-Based Systems*, 26:164 – 173, 2012.