Adam Standke
CSC 498-01

Literature Background Outline

1. Introduction
   - 1.1. Brief Description of Problem: Currently, many news based websites include content other than news articles. Often times, a news web page contains no articles at all but other forms of content such as videos, images, and/or news summaries. Given these different forms of content my research project aims to identify whether a news web page contains an article.
   - 1.2. Topics to be Covered: The selected topics to be covered include: 1) analyzing the link structure of a web page's URL; 2) using human perception to analyze a web page's structural layout; 3) analyzing a web page for content and visual features; and 4)machine learning through the use of C4.5.
   - 1.3. Relevance to Research Project: My research project is trying to solve the problem of identifying news articles. The first three selected topics help in identifying relevant features for the problem. Feature selection for the project is done using the two following methods: 1) observation (eg., analysis of URLs that contain news articles or through human perception);and 2) algorithmic extraction (eg., text-to-link ratio to find the tag that contains the most text or Maximum Scoring Subsequence that scores a web page based on continuous text sequences). The last selected topic uses a machine learning algorithm to learn a decision tree that can accurately predict whether a news web page contains an article .
2. Analyzing the link-structure of web page URLs
   - 2.1. Topic: Link-target analysis uses the structure of a link to classify the page as a news article or an irrelevant page. Every news web page contains a uniform resource locator (URL), which specifies the type of web page to be retrieved.
   - 2.2. Point 1: URLs that have an ID in their link structure are more likely to contain a news article
     - 2.2.1. "[T]he has number 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[5]."
   - 2.3. Point 2: URLs that have a date in their link structure are more likely to contain a news article
     - 2.3.1. "If the link contains a date, it is more likely to be a news article[5]."
   - 2.4. Point 3: URLs that have a longer link structure are more likely to contain a news article
     - 2.4.1. "Usually longer link structures are news[5]."
   - 2.5. Point 3: URLs that do not contain a certain type of reserved word are more likely to contain a news article

2.5.1. “[T]he has reserved word attribute was chosen to filter multimedia news. . . . [t]he 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[5].”

2.6. Point 4: URLs that contains a lot of slashes in the link structure are more likely to contain a news article

2.6.1. “Count the number of slashes in the link address, indicating the depth of the page in the website[5].”

2.7. Summary of Topic: A web page’s URL structure is an easy way to quickly identify whether a news web page might contain a news article. Analyzing the link structure of a URL before analyzing the actual HTML document provides a set of relevant features.

3. Using human perception to analyze a web page’s structural layout

3.1. Topic: Most websites structure/layout their content in ways that visually please human beings. This is especially true for news web pages, since they are designed to interact with humans. By analyzing how human readers scan news articles on the web, a structural layout can be constructed for web pages that contain news articles.

3.2. Point 1: An article on a news web page will usually have a specific layout structure

3.2.1. “News areas generally have some special properties that distinguish them from other areas. . . . 2) Space continuity: contents within a news area should be located continuously in space (vertically and/or horizontally), and generally they may only be separated by areas that are not pure text information (e.g., image areas, navigation areas, interaction areas, or decoration areas); 3) Formating continuity: major formats of different news areas should be similar[2].”

3.3. Point 2: A web page’s structure can be represented as blocks

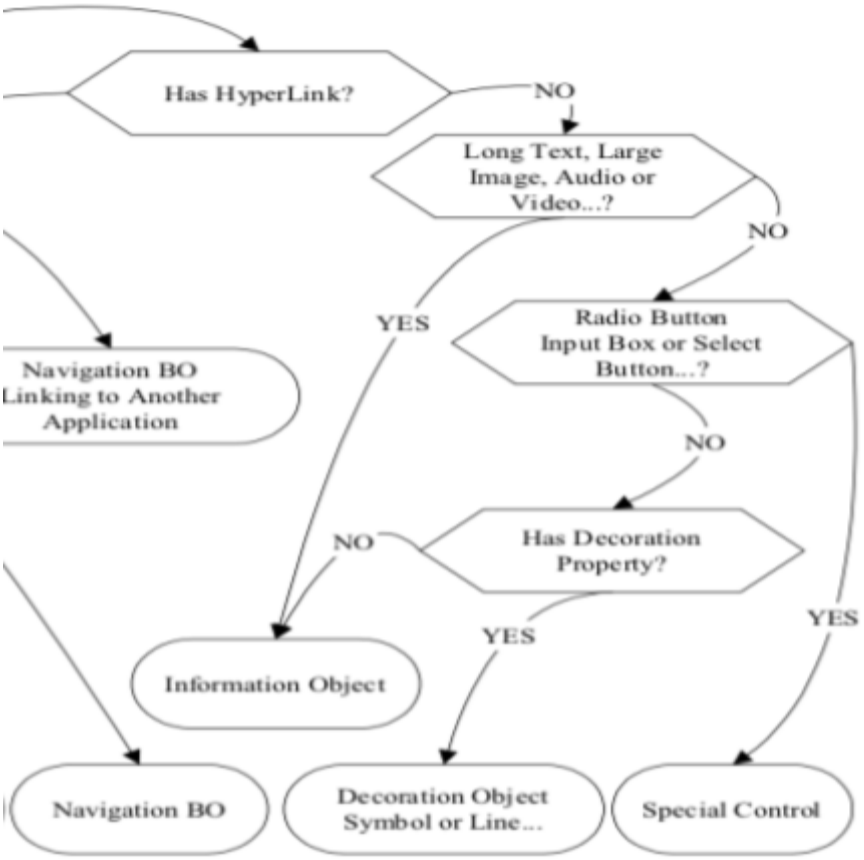
3.3.1. “[The VIPS algorithm] extract[s] the semantic structure for a web page. . . [the] semantic structure is a hierarchical structure in which each node will correspond to a block[1].”

3.3.2. “A web page. . . is a finite set of blocks. All these blocks are not overlapped. Each block can be recursively viewed as a sub-web page associated with sub-structure induced from the whole page structure[1].”

3.3.3. “[A] Web page is represented as an Object with certain functions. . . [a] Block Object is an Object displayed as independent content blocks separated from other blocks. . . [a] Leaf Block object is a Block Object that does not contain any other Block Object. . . [a] Leaf Block Information Object is a Leaf Block Object whose major function is to provide information[2]. ”

3.4. Point 3: A news article’s function is information and is thus classified as an information object

3.4.1. “[T]he function of a news content area is mainly to provide information[3].”



3.4.2. [3]

3.5. Point 4: A news web page that contains an article will be represented as a set of blocks containing information objects

3.5.1. "News content of a news Web page is presented as a set of. . . [Text Leaf Block Information Objects (TLBIOs)]. . . [n]ews TLBIOs of a news page are presented in one or more rectangular areas[2]."

```
Subroutine TLBIODetection (fp){
  TLBIOSet=∅;
  for each block level child c of fp{
    if c contains other Block Object
      if c is Information or Mixed Object
        TLBIODetection(c) →TLBIOSet;
      else if c' area ratio>ARThreshold
        TLBIODetection(c) →TLBIOSet;
    else if c is a Text Information Object c→TLBIOSet;
  }
  return TLBIOSet;
}
```

3.5.2. [2]

3.6. Summary of Topic: Most websites structure/layout their content in ways that visually please human beings. By leveraging human perception of online news, a structural layout can be assigned to areas of the web page that likely contain an article.

4. Analyzing a web page for content and visual features
    4.1. Topic: Content and visual feature analysis consists of examining the tags (ie., elements) and visual attributes (ie., style elements) of an HTML document. In doing so, a document object model (DOM) tree is constructed to analyze how the various tags and attributes relate to one another. Different types of algorithms and/or heuristics have been developed to identify the content and visual features of news articles.
    4.2. Point 1: An article on a news website can be described by certain HTML tags that usually deal with written text
        4.2.1. "After analyzing many HTML source files we find that a paragraph in a Web article can be created using different tags, including some paired-tags such as *div*, *p*, and *block-quote*, or some single-tags such as *br* and *hr* [6]."
    4.3. Point 2: From a visual and content perspective, an article on a news website consists of a continuous set of written text
        4.3.1. "For example, the article body usually consists of contiguous paragraph blocks occupying the main area of the [w]eb page[6]."
        4.3.2. "[B]y article we mean a contiguous, coherent work of prose on a single topic[7]."
        4.3.3. "Our approach is based on the observation that the article body usually consists of contiguous paragraph blocks occupying the main area of a web page[4]."
    4.4. Point 3: The written text contained by certain HTML tags in the DOM tree can be grouped into sequences of text segments
        4.4.1. "By the following steps we can group the text leaf nodes into multiple text segments[6]. "
        4.4.2. "Based on the visual attribute of line-break of DOM nodes, we identify the text segments that generally correspond to paragraphs[4]."
    4.5. Point 4: One way a news article can be located on a web page is by using an algorithm called Maximum Scoring Subsequence(MSS)
        4.5.1. "The consecutive sequence containing the main text body [ie., an article] is identified using the Maximum Scoring Subsequence (MSS) [algorithm][4]."
        4.5.2. "Typically, the text-body, a contiguous block of text. . . occupies the main area of the [w]eb page. Thus, we adopt the algorithm of MSS [ie., Maximum Scoring Subsequence] to identify the range of the text-body [ie., an article][6]."
        4.5.3. $$(a, b) = \arg\max_{x,y} \sum_{i=x}^{y} v_i$$

$$where\ v_i = F(s_i) \cdot StringLength(s_i)$$ [6]

    4.6. Point 5: A news article can also be located by finding the main content node

4.6.1. “CoreEx is motivated by the observation that the main content [ie., an article] in a news article page is contained in a DOM node. . . with significantly more text than links. For each node in the DOM tree, we pick a subset of its children which have a high text-to-link ratio. We then calculate the score of the node as a function of the total text and total number of links contained in this subset. The algorithm selects the node with the highest score[8].”

4.6.2.

- For a terminal node T :

  if T is a text node **then**
  $textCnt(T)$= word count of text in T, $linkCnt(T)$=0.
  else if T is a link node **then**
  $textCnt(T) = 1$, $linkCnt(T) = 1$.
  else
  $textCnt(T) = 0$, $linkCnt(T) = 0$.
  end if

- For a non-terminal node N:

  $textCnt(N) = 0$ and $linkCnt(N) = 0$
  S(N) is an empty set
  $setTextCnt(N) = 0$ and $setLinkCnt(N) = 0$
  for each child of N **do**
  $textCnt(N) = textCnt(N) + textCnt(child)$
  $linkCnt(N) = linkCnt(N) + linkCnt(child)$
  $childRatio = \frac{textCnt - linkCnt}{textCnt}$
  if $childRatio > Threshold$ **then**
  add child to S(N)
  $setTextCnt(N) = setTextCnt(N) + textCnt(child)$
  $setLinkCnt(N) = setLinkCnt(N) + linkCnt(child)$
  end if
  end for
  Store S(N), $textCnt(N)$, $linkCnt(N)$, $setTextCnt(N)$ and $setLinkCnt(N)$

4.7. Summary of Topic: Analysing a web page based on content and visual features allows an HTML web page to be broken down into its elemental form. This in turn allows patterns relating to how information is grouped within the document to be found. After finding these patterns different forms of information can then be extracted.

5. Machine learning through the use of C4.5

5.1. Topic: C4.5 is a machine learning algorithm that constructs classification models. Training data—comprised of attributes and classes are inputted into the algorithm. C4.5 will then generate a classifier in the form of a decision tree that acts as a model for predicting future outcomes.

5.2. Point 1: In constructing a decision tree C4.5 successively divides the set of training cases into subsets and proceeds until all of the subsets consist of cases belonging to a single class.

5.2.1.1. “[If] T [ie., the set of training examples] contains cases that belong to a mixture of classes. . . the idea is to refine T into subsets of cases that are, or seem to be heading towards, single-class collections of cases[9].”

5.3. Point 2: C4.5's test to partition the set of training examples is based on maximizing a criteria called *gain*

5.3.1.1. "A test is chosen, based on a single attribute[9]."

5.3.1.2. "The original ID3 used a criterion called gain. . . [t]he information theory that underpins this criterion can be given in one statement: [t]he information conveyed by a message depends on its probability and can be measured in bits as minus the logarithm to base 2 of that probability[9]."

5.3.1.3. "The quantity $gain(X) = info(T) - info_X(T)$ measures the information that is gained by partitioning T [ie., the set of training examples] in accordance with test X [ie.,an attribute].The gain criterion, then selects a test to maximize this information gain[9]."

5.4. Point 3: C4.5 can use a different criteria called the *gain ratio* to make more accurate predictions

5.4.1.1. "[T]he gain criterion. . . has a serious deficiency—it has a strong bias in favor of a test with many outcomes. . . [t]he bias inherent in the gain criteria can be rectified by a kind of normalization in which the apparent gain attributable to tests with many outcomes is adjusted[9]."

5.4.1.2. $gain\ ratio(X) = gain(X)/split\ info(X)$

5.4.1.3. "[T]he gain ratio criterion express the proportion of information generated by the split that is useful. . . [the gain ratio] selects a test to maximize the ratio. . . subject to the constraint that the information gain must be large[9]."

5.5. Point 4: C4.5 overcomes issues of overfitting by pruning a decision tree, resulting in a simpler and possibly even more accurate decision tree

5.5.1.1. "C4.5. . . now follows the second path [ie., removing subtrees retrospectively]. . . the overfitted tree [produced at first]. . .is then pruned[9]."

5.5.1.2. "[I]t is often possible to prune a decision tree so that it is both simpler and more accurate[9]."

5.6. Summary of Topic: C4.5 is a machine learning algorithm that constructs classification models. C4.5 constructs a decision tree from training data by recursively partitioning the data based upon a criteria called *gain*. C4.5 also provides additional features such as choosing a different partitioning criterion and the ability to prune a decision tree. Both of these additional features provide for more accurate predictions.

6. Conclusion

6.1. Summary of Topics: The topic of analyzing the structure of web page URLs has been shown to be beneficial in identifying web pages and their respective content.The topic of using human perception to deconstruct web pages provides a realistic framework for analysing an HTML document since web pages are constructed with human consumers in mind. The topic of analyzing a web page

for content and visual features takes an HTML document and transforms it into a semantic/hierarchical tree of nodes and leaves. Lastly, the topic of C4.5 deals with constructing accurate machine learning models that generalize well over unseen data.

6.2. How these topics support my project area: The first three topics help with identifying good attributes and values for my classification task. The last topic allows me to gauge whether the set of attributes used in the classification task are statistically significant.

Bibliography

[1] Cai, Deng, et al. “VIPS: a Vision-Based Page Segmentation Algorithm.” *Microsoft Technical Report (MSR-TR-2003-79)*, 2003.

[2]Chen, Jinlin, and Keli Xiao. “Perception-Oriented Online News Extraction.” *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL 08*, June 2008, pp. 363–367., doi:10.1145/1378889.1378952.

[3]Chen, Jinlin, et al. “Function-Based Object Model towards Website Adaptation.” *Proceedings of the Tenth International Conference on World Wide Web - WWW 01*, 2001, pp. 587–596., doi:10.1145/371920.372161.

[4]Fan, Jian, et al. “Article Clipper - A System for Web Article Extraction.” Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 11, 2011, doi:10.1145/2020408.2020525.

[5]Ferreira, Rodolfo, et al. “Appling Link Target Identification and Content Extraction to Improve Web News Summarization.” *Proceedings of the 2016 ACM Symposium on Document Engineering - DocEng 16*, Sept. 2016, pp. 197–200., doi:10.1145/2960811.2967158.

[6]Luo, Ping, et al. “Web Article Extraction for Web Printing: a DOM+Visual Based Approach.” *Proceedings of the 9th ACM Symposium on Document Engineering - DocEng 09*, Sept. 2009, doi:10.1145/1600193.1600208.

[7]Pasternack, Jeff, and Dan Roth. “Extracting Article Text from the Web with Maximum Subsequence Segmentation.” *Proceedings of the 18th International Conference on World Wide Web - WWW 09*, 2009, doi:10.1145/1526709.1526840.

[8]Prasad, Jyotika, and Andreas Paepcke. “Coreex.” *Proceeding of the 17th ACM Conference on Information and Knowledge Mining - CIKM 08*, Oct. 2008, doi:10.1145/1458082.1458295.

[9]Quinlan, John Ross. C4.5: Programs for Machine Learning. Morgan Kaufmann, 2006.