Adam Standke
Sr. Research 498

Annotated Bibliography

Citation:

Baykan, Eda, et al. “A Comprehensive Study of Features and Algorithms for URL-Based Topic Classification.” *ACM Transactions on the Web*, vol. 5, no. 3, 2011, pp. 1–29., doi:10.1145/1993053.1993057.

Description:

The paper deals with identifying the topic of a website by looking at its URL. The paper details a method of extracting features from a URL based on a token approach. Each URL is split into a sequence of strings and a bag-of-words is constructed that counts the number of tokens. Ultimately, the number of tokens that make up a URL are used as a feature in the topic classification process.

Source:

Journal Article

Relevant Aspects:

- A web page’s URL is broken down into a bunch of tokens that form a bag-of-words
- The number of tokens that form a web page’s URL can be used as a feature to determine whether a web page contains an article

Citation:

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.

Description:

The paper describes a way of extracting news articles from web pages by using a combination of link identification and content extraction. The paper details six features that help in determining whether links on a web page deal with a news article or irrelevant subject matter. In addition to classifying links the paper also outlines a content extraction method. Namely, the content in each tag is replaced by the number of words in the tag and a tree is constructed to find the area of a webpage with the most text.

Source:

Conference Proceeding

Relevant Aspects:

- The attributes that make up web page links are thoroughly detailed
- Determining whether a URL contains a date and how many slashes make up a URL can be used as additional features

Citation:

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.

Description:

The paper demonstrates a system that automatically extracts news content through the use of text extraction, title extraction, image and caption extraction, and next-page link detection. Using a heuristic approach that the article body usually consists of contiguous paragraph blocks, an

article is detected using an algorithm called Maximum Scoring Subsequence (MSS). MSS identifies the text segments that correspond to paragraph tags and creates a sequence of text segments from the DOM. The goal is to find a sequence of text segments that bounds a textural region. Also, the paper details a title extraction method that determines whether a web page contains a news title by using various criteria such as horizontal starting position, font size, and distance from a title tag.

Source:

Journal Article

Relevant Aspects:

- Maximum Scoring Subsequence(MSS) identifies text segments that correspond to paragraph tags in the DOM and determines the largest contiguous sequence of text segments in a web page
- The length of this maximum subsequence can be used as an additional feature to determine whether a news web page contains an article

Citation:

Wang, Junfeng, et al. "Can We Learn a Template-Independent Wrapper for News Article Extraction from a Single Training Site?" *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 09*, 2009, doi:10.1145/1557019.1557163.

Description:

The paper tackles the problem of web page news extraction without the use of templates. The paper uses the DOM to construct subtrees for the title and article portions of a web page. By

using both spatial and content features of a web page an article extraction algorithm is proposed. This extraction algorithm is composed of two parts. The first part computes the most likely area of a web page that may contain a news article. The second part computes a list of candidate areas that may contain an article section. Also, the paper leverages the high correlation between a news title and article to extract news titles.

Source:

Conference Proceeding

Relevant Aspects:

- Using both content, spatial features, and general observations about news articles an algorithm for generating possible areas where a news article may reside is presented
- Will use this algorithm to generate a number of candidate areas
- The number of possible areas where a news article may reside can be used as an additional feature to determine whether a news web page contains an article

Citation:

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.

Description:

The paper developed an algorithm for extracting the main article from a news website. The paper constructs a DOM tree of a web page and scores every node based on a text-to-link ratio. The algorithm selects the node with the highest score as the main content node and stores the set of nodes that actually contain the article in a separate data structure. The paper also details various heuristics and preprocessing steps to use when extracting news articles.

Source:

Conference Proceeding

Relevant Aspects:

- Details an algorithm for determining if a web page contains an article by using the heuristic that the main content in a news article will have significantly more text than links
- Will use this algorithm to find the main content node in a DOM
- The set of nodes that are the children of the main content node can be used as an additional feature to determine whether a news web page contains an article

Citation:

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.

Description:

The paper discusses a way to extract an article from a news website by implementing different forms of the Maximum Subsequence Segmentation (MSS) algorithm. MSS uses local, token level classifiers to find a score for each token in the document resulting in a sequence of scores. The paper outlines a baseline implementation of MSS called Simple MSS that the paper just used for benchmarking purposes. In Simple MSS a maximum subsequence is found by generating a sequence of scores and assigning a score of -3.25 to a tag in the DOM and a value of 1 to every word and symbol contained within the tag. This simple approach to MSS produced a 90% precision rate for extracting the main article of a news website.

Source:

Conference Proceeding

Relevant Aspects:

- Details an algorithm called Simple MSS that finds a maximum subsequence by assigning a score of -3.25 to a tag in the DOM and a value of 1 to every word and symbol contained in the tag
- The subsequence with the highest score is assumed to be an article in a news web page.
- The length of this subsequence can be used as an additional feature to determine whether a news web page contains an article

Citation:

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.

Description:

The paper describes an online news extraction approach based on human perception by detecting areas where an article may reside based on content functions, space continuity, and formatting continuity of news information. The paper develops a function based object model in which each piece of content in a web page is represented as an object that serves a certain purpose. The paper outlines a set of axioms to simulate human perception of news extraction. Based on these axioms the paper develops an algorithm that seeks to find sets of Text Leaf Block Information Objects (TLBIOs). These TLBIOs are presented in one or more rectangular areas of a web page and form possible areas in which a news article may reside.

Source:

Conference Proceeding

Relevant Aspects:

- Details a perception based approach to news extraction based on how humans identify online news
- Creates a Function Based Object Model (FOM) that transforms a web page into a set of objects, such as information objects, navigation objects, interaction objects and decoration objects
- Provides axioms to create a perception-oriented news extraction algorithm that takes as input a news web page and creates a set of Text Leaf Block Information Objects (TLBIOs), which are later used to detect possible areas in which a news article may reside.
- Detection of areas in which a news article may reside can be used as an additional feature

Citation:

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.

Description:

The paper presents a function-based object model that attempts to identify relationships between objects in a web page. In a function-based object model objects can be classified into basic objects (BOs) or composite objects (COs) that have either general or specific functions. The

paper outlines a decision tree method for transforming BOs into different types of function objects, such as information objects, navigation objects, and decoration objects.

Source:

Conference Proceeding

Relevant Aspects:

- Outlines steps that transform HTML tags into basic objects (BOs) such as information objects, navigation objects, interaction objects, and decoration objects.
- The Function Object Model Analysis presented in the paper is a principal component of the perception-oriented news extraction algorithm

Citation:

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

Description:

The paper details an algorithm called VIPS that extracts the content structure of a web page by segmenting the DOM tree. The segmentation process has three steps. First, the DOM tree is broken down into blocks based on a set of heuristics. Second, the blocks are separated horizontally or vertically by the weighing of visual separators. And lastly, the content structure of the page is established by merging blocks together based on separator weights.

Source:

Technical Report

Relevant Aspects:

- The VIPS algorithm detects the content structure/layout of a web page

- The VIPS algorithm could be used to see if web pages that contain news articles have a different content structure than web pages that do not contain news articles

Citation:

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.

Description:

The paper proposes a novel method of article extraction using both the DOM and visual features. The paper details a web page segmentation algorithm that identifies line-breaking features in a web page. These line-breaking nodes are usually found in paired tags such as div, p, br, and hr or in the CSS display property.The text segments are grouped together and then Maximum Scoring Subsequence(MSS) is applied to identify the maximum subsequence bounding a text-body.

Source:

Journal Article

Relevant Aspects:

- Gives additional heuristics for Maximum Scoring Subsequence(MSS); an algorithm that determines the largest contiguous sequence of text segments in a web page
- The length of this maximum subsequence can be used as an additional feature to determine whether a news web page contains an article

Citation:

Quinlan, John Ross. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 2006.

Description:

The book details the algorithms that underpin C4.5 through a series of working examples. The book serves as a manual for using C4.5. Chapters 2 through 8 detail how to create decision trees and generalize rules from them. And chapter 9 explains how to run and conduct experiments using C4.5.

Source:

Book

Relevant Aspects:

- C4.5 will serve as the machine learning algorithm for my research project
- A decision tree will be constructed using features specifically tailored for news websites.
- Each training instance will be a rendered DOM tree of a news web page and manually labeled as either containing a news article or not
- C4.5 will decide if a news web page contains an article

Citation:

Roldán, Juan C., et al. "Extracting Web Information Using Representation Patterns." *Proceedings of the Fifth ACM/IEEE Workshop on Hot Topics in Web Systems and Technologies - HotWeb 17*, 14 Oct. 2017, doi:10.1145/3132465.3133840.

Description:

The paper presents an information extraction technique based on web representation patterns. The paper's information extraction technique is domain independent. Furthermore, it takes an unsupervised learning approach towards extracting web information based on different web heuristics. Even though the paper does not specifically focus on the extraction of news articles

it does provide some preliminary concepts and heuristics that are useful for pattern extraction and analysis of DOM trees.

Source:

Conference Proceeding

Relevant Aspects:

- Provides heuristics for extracting different types of information found on web pages
- Heuristics can be applied when analyzing the DOM tree
- Extraction techniques are based on a rendered DOM— not just the underlying source code

Irrelevant References from Preliminary Bibliography

Citation:

Krüpl-Sypien, Bernhard, et al. "A Versatile Model for Web Page Representation, Information Extraction and Content Re-Packaging." *Proceedings of the 11th ACM Symposium on Document Engineering - DocEng 11*, 2011, doi:10.1145/2034691.2034721

Source:

Conference Proceeding

Description:

While the article did contain information related to web extraction, the article was not chosen for my annotated bibliography for a couple of reasons. First, the article did not give the type of implementation details needed to recreate the Unified Ontological Model(UOM). The article was more abstract and relied on theories that would have taken additional research to understand. Namely, gestalt theory which is a psychological theory that tries to understand how humans group objects. Also, the article's UOM would have required examining browser calls to a graphical unit. Also, the paper did not focus as much on the DOM in constructing its UOM model.

Citation:

Freund, Luanne, et al. "Digging into Digg." *Proceedings of the 2011 IConference on - IConference 11*, 2011, doi:10.1145/1940761.1940864.

Source:

Conference Proceeding

Description:

The article did not relate to my research topic since it focused on characterizing common forms of online news content. The article's work was primarily focused on developing a taxonomy of common textual news genres. To do so, the article focused on a website called Digg to get a general understanding of the cross-section of different types of news content. The article is irrelevant for my research project since my classification problem does not deal with understanding different types of news content.

Citation:

Arasu, Arvind, and H. Garcia-Molina. "Extracting Structured Data from Web Pages (Poster)." *Proceedings 19th International Conference on Data Engineering (Cat. No.03CH37405)*, doi:10.1109/icde.2003.1260839.

Source:

Conference Proceeding

Description:

While the article was very informative and did provide implementation details for information extraction, the article's primary focus was extracting structured data (ie., tabular data found on sites like Amazon or ebay). Web pages are in general semi-structured and harder to extract information from. Since news web pages are generally not structured in a tabular form this article is irrelevant for my research project.

Citation:

Quinlan, J.r. "Induction of Decision Trees." *Machine Learning*, vol. 1, no. 1, 1986, pp. 81–106., doi:10.1023/a:1022643204877.

Source:

Journal Article

Description:

The article deals with the ID3 machine learning algorithm which is the precursor to the machine learning algorithm C4.5. For my project I will be using C4.5 as my machine learning algorithm. I have a book that uses and implements the C4.5 algorithm. Hence, there is no reason for me to keep this article when C4.5 includes additional implementation details that ID3 does not have.

Citation:

Baykan, Eda, et al. “Purely URL-Based Topic Classification.” *Proceedings of the 18th International Conference on World Wide Web - WWW 09*, 2009, doi:10.1145/1526709.1526880.

Source:

Conference Proceeding

Description:

The article did not provide the type of in-depth explanation of URL-based topic classification. Also, the article did not provide another type of token based approach other than n-gram analysis. Since I do not want to take an n-gram approach, the article's experimental results would be irrelevant for my research project.

Citation:

Krüpl, Bernhard, et al. “Using Visual Cues for Extraction of Tabular Data from Arbitrary HTML Documents.” *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web - WWW 05*, 2005, doi:10.1145/1062745.1062838.

Source:

Conference Proceeding

Description:

The article's primary focus was extracting structured data (ie., tabular data found on sites like Amazon or ebay). Web pages are in general semi-structured and harder to extract information from. Since news web pages are generally not structured in a tabular form this article is irrelevant for my research project.

Citation:

Jirkovsky, Vojtech, and Ivan Jelinek. “Method Combination for Information Extraction.” *International Conference on Computer Systems and Technologies - CompSysTech’10*, 2010.

Source:

Conference Proceeding

Description:

The article did not provide any implementation details but rather described the three main methods used today to extract information from the web. The article talked about three methods of web page extraction.The article did not detail how these various methods could be implemented or applied. The article was more of a summary of the type of methods that are being used today to extract information from the web. Also, this article provided no performance metrics. Thus, the article is irrelevant for my research project.

Citation:

Jiménez, Patricia, and Rafael Corchuelo. “Roller: a Novel Approach to Web Information Extraction.” *Knowledge and Information Systems*, vol. 49, no. 1, 2016, pp. 197–241., doi:10.1007/s10115-016-0921-4.

Source:

Journal Article

Description:

The article details a system called Roller. Roller learns rules that are then used to extract information from semi-structured web documents. When Roller is given a document that has been annotated in some way, Roller will find the best rule that extracts the information. This is done using a combination of attributes and relational features. Also, the article’s primary focus is on creating rules for information extraction. My research project is focused on classifying web pages; thus, the article is irrelevant since Roller is a rule based learner.

Critique of Automatic Citation Tool

I used the automatic citation tool called bibme.org. Each of the citations contained in my annotated bibliography and list of irrelevent references was done using MLA format. While the site is free to use and does a decent job in finding, extracting, and citing articles two things did bother me when I used the site. First, the site did not have a reference to every article typed into the citation search bar. Consequently, I had to fill in missing information such as the date and title. Second, the site is free but does contain a lot of advertisements which can become distracting after a while. In certain circumstances the site makes you watch an advertisement and makes you rate it before being able to continue using the site's services. Other than these minor issues I would recommend this automatic citation tool since it makes the citation process very easy.