A Prescriptive Approach For Structured Information Extraction From Web Forums And Social Media

CUMBERLAND, Ethan and DAY, Tony <http://orcid.org/0000-0002-3214-6667>

A Prescriptive Approach For Structured Information Extraction From Web Forums And Social Media

Ethan Cumberland
CENTRIC
Sheffield Hallam University
Sheffield, England
e.cumberland@shu.ac.uk

Tony Day
CENTRIC
Sheffield Hallam University
Sheffield, England
t.day@shu.ac.uk

ABSTRACT

In this paper we present ongoing research into extracting highly structured data - such as authors, posts, the links between them, and the metadata about them - from social media and fora using a prescriptive approach, building upon simple observations and generalised rules. This method uses techniques designed around identifying content based on text features, such as text density, and combines it with simple rules derived from studying the common structures of the target web pages to infer and extract structure from structured data.

We discuss observations made from studying a number of social media websites and forums and present the simple rules for post, content and attribute identification developed from these observations. We also present the structured format used to store the extracted data and some of the benefits of this structure. Next, we give initial experimental results, showing that the proposed approach can achieve accuracies above 90% for identifying posts, 70% for extracting content from these posts, and 50-70% for extracting additional attributes about the posts and their authors. We highlight factors influencing these results, before finally detailing the next steps for this research.

Our research shows that it is possible to achieve reasonable levels of accuracy for extracting structured data using an approach that requires no training and is transferable between different social media and web forums with no additional input necessary. This approach thus promises considerable efficiency gains compared to the training involved with current machine learning-based approaches, whilst maintaining reasonable performance.

CCS Concepts

• Information systems → Information retrieval; Web mining

Keywords

web scraping, web forums, social media, structured information extraction, wrapper induction

This work was carried out under the CONNEXIONs project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 786731 (CONNEXIONs – InterCONnected NEXt-Generation Immersive IoT Platform of Crime and Terrorism DetectiON, PredictiON, InvestigatiON, and PreventiON Services). The information in this paper reflects only the authors' view and the Agency is not responsible for any use that may be made of the information it contains.

1. INTRODUCTION

The web is a treasure trove of information that is easy to see but difficult to acquire and use. The growth of the web has seen a massive rise in the number of users of social media and web forums, with Facebook reporting an increase from 618 million average daily active users (DAUs) in December 2012 [4] to 1.84 billion DAUs in December 2020 [5], and popular discussion website Reddit reporting 52 million DAUs in October 2020, an increase of 44% year-on-year [10]. This has led to a huge increase in user-generated content that is available online. For example, Reddit reported that in 2020, 303.4 million posts and 2 billion comments were made on the site, up 52.4% and 18.6% respectively year-on-year [10].

With this abundance of potentially useful information available on the internet, systems for extracting information are required to be able to retrieve some of this data to allow it to be used. Anyone wishing to use the data could always extract it manually, but given the scale of the data, that immediately becomes implausible. There are many reasons people might want to use data from social media, forums and the web in general, from forming the basis for technologies like automatic summarization [12] to collecting marketing data about companies posted by users online [9].

Extracting the textual content from these online sources does not have to be the end of the road, however. The data that is available on social media websites and forums stems beyond just textual content, it includes information about the users that posted that content as well as additional attributes about these posts such as the post time, likes and shares of the posts, and much more. Finding a way to extract all of this structured data and store it in a similarly structured way would increase the potential applications of that newly extracted data, allowing for more powerful tools to be developed with it, and more powerful operations to take place using it.

Extracting structured data from the web is a long-standing area of research with a wide variety of approaches proposed for tackling the task from a multitude of angles and for various reasons [1, 7, 11, 14]. Many proposed solutions to this problem involve complex machine learning techniques from visual-based content detection [3] to unsupervised learning approaches [14].

In this paper we briefly describe the problem domain and some existing approaches, before going on to propose a

prescriptive, rule-based approach to extracting content from social media websites and forums and storing that data in a structured format to enrich this data, evaluating the initial performance of the rules defined during our research, and drawing some conclusions and steps for the future direction of this research.

We focus on the technical implementation of these rules within this paper, but the replication, implementation and usage of anything mentioned here should be carried out within the appropriate legal framework, with ethical oversight, and in accordance with all rules, regulations, terms and conditions of any websites that this is practiced upon.

2. RELATED WORK

The topic of extracting data from the web has been an active area of research for many years and has been tacked from numerous angles. Earlier research into content extraction often focussed on extracting main content from online sources such as news articles. Some of these early methods include extracting the main text by using the ratio of tags within HTML documents [13] and the density of text within HTML documents [6], among many other.

Building upon this work are approaches designed to extract more structured data from sources containing more structured information. These sources include web forums, social media websites, online marketplaces, and other web pages that often contain multiple blocks of data that are independent from each other, each with their own content and attributes, such as forum posts and product listings. The task of extracting data from these sources is different to general web or main content extraction as not only are the types of web pages often vastly different from those containing single pieces of content such as news articles, but the methods used to extract this data must now be able to identify and perform extraction on each piece of content individually, accounting for variation between pages on the site.

Many of the approaches to solving this problem attempt to extract data using the structure of the DOM tree of HTML pages, using this often-repetitive structure to create wrappers and identify templates from which to extract the desired information. These template-dependent approaches [8, 15] - approaches that work based on the repetitive structure of HTML pages - often use machine learning-based methods to learn page structures and enable the data extraction to take place. The alternative is template-independent approaches, approaches that are designed to work across multiple websites where the structure of the pages are different, often employing probabilistic methods to identify the structure of the individual pages they are set to work on [1, 7, 11, 14].

One of the biggest issues with many current approaches, in particular the machine-learning based approaches, is that the models that power them require training before they can be used, and often require training on large set of training data relevant to the domain for which extraction is intended to be performed on. This training data can be hard to come by, especially in the age of increased information security, privacy regulations, and other strict terms of services which govern the platforms from which ideal training data could be collected from.

3. THE PROPOSED APPROACH

The proposed approach is based on using simple rules which have been defined by investigating the structure of some social media websites and forums, such as the fact that content within posts usually sits in the same position relative to each post. These rules can then be used to create models of websites that point to the posts and content within them, and these models can be used to extract the content and store them in a structured format for further use. The technical details of these models is discussed in Section 3.1, and the structured format that the data is extracted to is discussed in Section 3.5.

The dataset of social media websites and forums chosen to analyse for this consists of the forums of the 10 most popular forum software technologies' websites in the top 1 million sites by traffic [2], plus the popular social media and discussion sites Facebook, Twitter, Reddit, VK and Gab. The forums of popular forum software technologies were chosen because while the websites that use them can often use their own themes on top of the forum software to change how they look visually, the underlying HTML and page structure is often the same, so by using these as a basis we can encompass many forum websites by analysing only a few.

Despite their popularity, image-based social media sites such as Instagram and video-based sites such as TikTok were excluded from this as the method of content identification chosen is based on textual content, so these websites would not be suitable for the types of identification we aimed to perform in this paper.

While this set of forums and social media websites is not representative of all forums and social media websites on the internet, and excludes custom and bespoke forums, choosing some of the most popular should mean that the output of this research is functional and applicable to as many situations as possible as quickly as possible. Section 5 details the steps that will be taken to improve the work presented in this paper and expand its functionality.

3.1 Technical Implementation

In order to both evaluate and use the rules we defined, we chose to implement the approach in a Java application to perform the identification and extraction. This Java application was fed pages from a browser rather than fetching them directly to ensure that any client-side code was run, and any data fetched via requests was acquired prior to analysing and extracting. It also features a database to store the models that are generated so that they can be reused.

The application takes the HTML of given forums and social media websites as input, performs some pre-processing – removing HTML elements that only provide metadata but no content: `<head>`, `<link>`, `<meta>`, `<script>`, `<noscript>`, and `<style>` - and then applies the rules to create a model for that website. This model consists of the domain of the forum or social media website that the model represents and a collection of tag paths that represent the location of posts, content, and attributes on the page.

"A tag path is, similar to an XPath expression, a sequence of HTML tags, that corresponds to the path from the

HTML root node to the respective HTML element from which the block segment is built" [11]. These tag paths were used to represent the path to posts within the pages, as well as the relative path from posts to the content and attributes within posts.

With regards to usage of the approach we propose, there are two possibilities:

1. Creating a model for each page that is visited and using that model to extract information from that page only. This is using the techniques in a template-independent way, the model is based on that site only and does not benefit from any information outside of what was acquired from that single page.
2. Creating a model for a single type of page on a website and using that model to extract information across the entire site. This is using the techniques in a template-dependent way, the model can be used across the website without needing a new one to be created for every page.

We believe that the second method of using the approach we developed is the better method, as not only does it mean less models have to be created to extract the same amount of data, but these models are also more resistant to change between each page provided that the model was generated on a page on the site that is representative of the rest of the site. Generating a model on a content-rich page which is representative of the content and its variation would create a more robust model than one generated on a page with a low amount of content, and this model would then be applicable to pages with both large and small amounts of content.

The following sections will discuss the investigations that lead to the definition of the rules that make up this approach.

3.2 Post and Content Identification

The first step to being able to identify and extract any amount of information from the posts on these pages is to first identify the posts and the content within them. When looking at the structure of social media websites and forums we see a lot of repetition in the structure of the DOM tree.

Figure 1 shows some mock forum posts, and Figure 2 shows those mock post elements within the DOM tree of the page, showcasing how the posts are all represented by consecutive elements within the HTML.

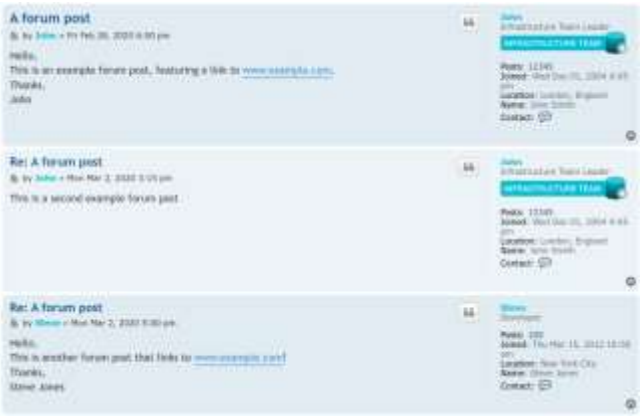

Figure 1: Mock forum posts from the phpBB forums.

Figure 2: The HTML for the mock phpBB forum posts in Figure 1 with the three posts highlighted.

However, if we were to look for only repetitive elements with similar structures within the page, we would identify a lot of false positives, such as the navigation elements on the navigation bar of the phpBB forum as shown in Figure 3.

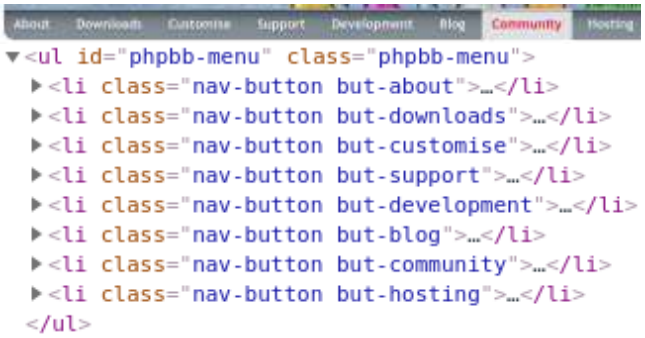

Figure 3: phpBB forum's navigation bar and its corresponding HTML.

Because of this, we need to look deeper into the post elements to identify something we can use to distinguish between them and other repetitive elements on the page, such as the content within the posts.

To identify content in the page, and posts based on this, we build upon the idea of using text density to identify content blocks as proposed by Sun et al. [6]. Sun et al. defined the text density of a node n, TD_n, as the ratio of the number of characters within and under the node n, C_n, to the number of tags under that node, T_n.

$$TD_n = \frac{C_n}{T_n} \quad (1)$$

After calculating the text density of the elements in the DOM tree, we look for the highest density elements and calculate the tag paths for these. We then looked for which of these high-density elements were repeated many times and which of them had similar tag paths that would indicate them being at the same position relative to each post. Using these elements, and looking for a common parent element they share, we were able to identify posts and the content elements within the page, which led to the first two rules:

Rule #1: Social media websites and forums are often structured so that there is one element on the page that contains all post elements as direct children at the same depth in the DOM tree, and as such all posts can be identified with a single tag path.

Rule #2: The content within these posts are usually the largest elements within each post in terms of text density when considering the average of all posts on the page and are often in the same position in the DOM tree relative to each post. The content within each post can therefore be identified with a single tag path.

3.3 Author Identification

When it comes to author identification, we are looking to identify the author's username or display name, profile URL and profile photo URL, where available. Despite how it appears visually on the page, in the DOM tree of every one of the websites within the set of social media websites and forums, the author's username, profile URL and profile photo appear before the main content block within each post, if they are present at all. When looking at these usernames/display names and profile URLs, we identified that every website in our dataset represented the username/display name as a `<a>` element with a `href` attribute containing the user's profile URL.

Rule #3: Author usernames are often `<a>` elements that link to the user's profile on the social media website or forum.

While identifying the author's usernames/display names and profile URLs, it became apparent that information about the author of each post is often grouped together under one HTML element, and this element is usually a sibling to the element that contains the content of each post. This can include other information such as number of posts, forum ranks, the date of account creation, etc. While we haven't attempted to identify these additional attributes during our research so far, the grouping of these elements allows us to define another rule regarding the location of the post author's information:

Rule #4: When looking at the structure of a post, the element containing information about the author of the post often comes before the content of the post, and these author and content elements are often direct siblings under some parent element.

Now that we have attempted to identify the author's usernames and profile URLs, we attempted to identify their profile photos where available, which was on 12 of the 15 websites in our dataset. Profile photos, on the websites that allow them, are often displayed near the user's username and other profile details. While this gave us a good indication on how to pick them out, there were more images that were close to these elements such as badges, status indicators, and other icons like these. To be able to distinguish between them, we looked at how all of these other icons and badges are represented in the HTML, and it became clear that profile photos are often represented by `<image>` and `<img>` elements, whereas these other icons were all `<i>` elements and SVGs.

This distinction allowed us to identify our 5th rule:

Rule #5: Author profile photos are often the closest image element in the DOM tree to the author's username.

3.4 Attribute Identification

Attribute identification is something we have yet to properly investigate, however, to show that it is at least feasible, we attempted to identify the created time of the posts. Thanks to the adoption of newer HTML5 elements, a majority of the sites within our dataset attribute post time using a `<time>` element, which is easily identifiable. However, this is not the case for all sites in our dataset, and as such all social media websites and forums. The sites in our dataset that did not opt to use the `<time>` element instead provided the time as `datetime` attributes on elements, or simply as text within `<span>` and `<p>` elements with datetime-related class names such as `date`, `timestamp` and `timeago`.

Facebook is a notable exception to this, as the post time element is not uniquely identifiable outside of generated class names that are not human-readable or easily understood.

While using this set of element and attribute names is not the most generic approach that could be taken, this is ongoing research and therefore there is scope to improve this in the future, to make it more generalised and attempt to capture some of these edge cases. This, therefore, led to the definition of our 6th, and loosest, rule:

Rule #6: Timestamps are usually represented as one of only a few formats, including `<time>` elements and other elements such as `<span>` and `<p>` elements with classes such as `date` and `timestamp` and attributes such as `timestamp`.

3.5 Structured Data Model

Storing the extracted data in a structured format is key to adding value to that data beyond just simply extracting it and storing it verbatim. Here we show the structured format we used for storing the data we extracted, with an example, and discuss how it can enrich the data that is stored within it.

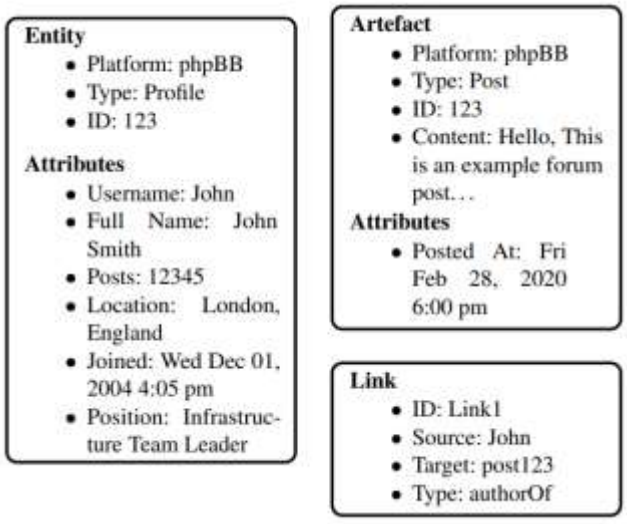

Figure 4: An example of a post and the information it contains stored in the structured data model.

The structured data model consists of three types of data: Artefacts, Entities and Links. Artefacts are used to represent pieces of content and the attributes about them, Entities are used

to represent things such as people, objects, locations and events, and Links are used to represent the relationships between Artefacts and Entities. Figure 4 shows what the first post in Figure 1 would look like when stored in this data model once all content and metadata have been extracted, excluding the link to the external URL.

While these three types of data alone are flat data structures, using artefacts, entities and links allows us to achieve a graph-style representation of this data, and a high degree of linking between them. Figure 5 shows what the forum posts shown in Figure 1 could look like when stored in the structured data model and viewed in a graph-style representation.

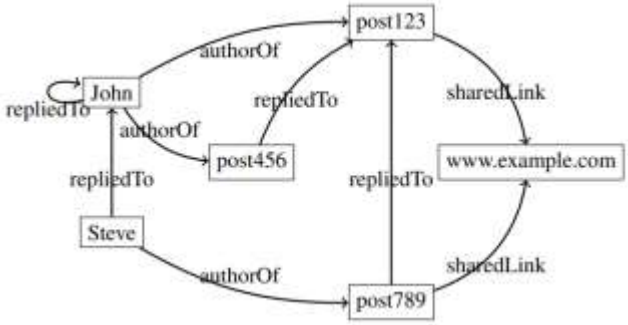

Figure 5: An example of the graph-style representation possible using the structured data model.

As shown by the multiple links between the entity 'John' and the other elements on the graph in Figure 5, this data structure would allow for all posts made by one author to link back to the same author entity within the data structure by using the identifiers from the forum/social media website when creating links, to ensure that there is no duplication of entities and artefacts. This would allow for profiles extracted into this data model to be viewed as a network graph, enabling analysis like Social Network Analysis to take place on the data, which would not be possible if the forum or social media content was extracted and stored as simple text.

4. EVALUATION

To assess the performance of the approach at this early stage, an evaluation exercise was performed on 50 pages each from 4 sample sites. This sample set includes pages from Reddit, a site that was analysed when developing the approach, Gentoo Forums and The Tech Report Forums, both built upon phpBB which was also analysed in the development process, and Microsoft's ASP.NET Forums. While this sample set is small, this evaluation exists to show that this prescriptive approach can begin to identify data across sites that vary in structure, prove the flexibility of the approach even after only a short period of research and development, and to provide a baseline for further improvements to build upon.

The accuracy of the technical implementation of the rules we defined was calculated for both of the possible methods of use outlined in Section 3.1. The accuracy of the identified posts is equal to the number of posts collected, P, over the number of posts on the page, P_{max}:

$$Accuracy = \frac{P}{P_{max}} \quad (2)$$

The accuracy of the content, timestamp, username, profile URL and profile photo URL identification was then based on the number of correctly identified posts as opposed to the number of the posts on the page, meaning that an accuracy of 0.500 for content identification would signify that the content was successfully identified in half of the posts that were successfully identified on the page.

Table 1: Accuracy of extraction using a model created on each page of a website

Website	Posts	Content	Timestamps	Usernames	Profile URLs	Profile Photo URLs
Reddit	0.920	0.380	0.709	0.709	0.709	-[1]
ASP.NET Forums	0.937	0.838	1.000	1.000	1.000	0.937
Gentoo Forums	0.923	0.966	0.000	0.000	0.000	0.000
The Tech Report Forums	0.898	1.000	1.000	1.000	1.000	-[1]
Average[2]	**0.915**	**0.713**	**0.684**	**0.684**	**0.678**	**0.502**

Table 2: Accuracy of extraction using a model created on one page of each website and used across all pages of that same website.

Website	Posts	Content	Timestamps	Usernames	Profile URLs	Profile Photo URLs
Reddit	1.000	0.222	0.424	0.424	0.424	-
ASP.NET Forums	1.000	1.000	1.000	1.000	1.000	1.000
Gentoo Forums	1.000	1..000	0.000	0.000	0.000	0.000
The Tech Report Forums	1.000	1.000	1.000	1.000	1.000	-[1]
Average[2]	**1.000**	**0.688**	**0.602**	**0.684**	**0.678**	**0.536**

[1]This site either does not have user profile photos or does not show user profile photos on the post pages
[2]The number of posts varies across each site, so these averages are weighted to account for this

4.1 Experimental Results

For an initial test of the rules we defined, the post and content identification performed better than expected, with username, profile URL and profile photo URL identification performing almost perfectly on some sites, but not so well or not at all on others.

As a general trend, the results in Tables 1 and 2 show two things:

1. When the model for a website is identified on a single page and used across the website such as in Table 2, so long as the model is representative of the rest of the pages on that website, the model will perform well thanks to the regularity of the pages on the site.
2. The rules that have been developed are good for identifying posts and content but are not yet generalised enough for reliable identification of additional attributes in all situations. This can be seen in the results of attempting extraction from Gentoo Forums. It is worth noting that on all pages on Gentoo Forums where content was identified from posts, the parent DOM element containing the author attributes was identified, but the attributes themselves were not.

The poor performance of this approach on Reddit can be attributed to the fact that in multiple posts on the same page, the location of content and attributes varies depending on the type of post - some posts may contain images as the main content, others may contain varying amounts of text content, and some are advertisements styled as posts - each of which is rendered differently on the page. When identifying the model for any given page, the model will be generated for whichever of these types appears most frequently, hence the disparity between the accuracy on Reddit and the other websites.

5. CONCLUSIONS AND FURTHER WORK

In this paper we have presented ongoing research into extracting highly structured data from social media and fora using a prescriptive approach. Despite the shortcomings we identified from analysing the results in Section 4, the performance of this approach on sites like ASP.NET Forums and The Tech Report Forums is promising.

We have shown that by analysing the structure of content on a number of social media websites and forums, it is possible to draw enough parallels between them to create generalised rules that can work on unseen websites with good accuracy. The fact that this has been achieved, with accuracies of up to 100% on ASP.NET Forums and The Tech Report Forums, without needing training datasets or any model training at all, means that data extraction of this type can become more accessible.

We also presented a structured model for storing the data that can be captured using this approach to further enrich the extracted data and provide more value for potential uses of data extracted in this format.

The approach we proposed is currently limited by the scope of our investigatory work performed when identifying the rules we proposed, which is why the next step for validating and improving the work done and the rules defined in this paper is to acquire more data. Firstly, we need to identify a larger dataset of sites for helping to establish, generalise, expand upon, and improve the rules defined in this paper to hopefully lead to improved performance and a more robust set of rules ready to tackle the problem. In addition to a larger testing dataset, we need a larger, more representative dataset of social media websites and forums to evaluate the rules' performance.

Once a large enough testing dataset has been identified, we can consider comparing the approach developed during our research with existing approaches and software solutions to see the strengths and weaknesses of each for the task at hand and identify areas in which our approach is lacking compared to existing solutions and can be improved upon.

Another area where improvements could be made to this approach lies around the technical implementation of what we proposed in this paper. If the approach was to be implemented as a software tool, it would enable us to add features such as making edits or corrections to the models generated using the rules and storing and sharing these models. While these additions are out of scope for the current research activities, they could improve the feasibility of using the approach we define in terms of real-world usage beyond the research.

REFERENCES

[1] Bill Yuchen Lin, Ying Sheng, Nguyen Vo, and Sandeep Tata. 2020. FreeDOM: A Transferable Neural Architecture for Structured Information Extraction on Web Documents. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 1092–1102. https://doi.org/10.1145/3394486.3403153

[2] BuiltWith® Pty Ltd. Forum Software Usage Distribution in the Top 1 Million Sites. https://trends.builtwith.com/cms/forum-software

[3] Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. 2003. *VIPS: a Vision-based Page Segmentation Algorithm*. Technical Report MSR-TR-2003- 79. Microsoft Research. 28 pages. https://www.microsoft.com/en-us/research/publication/vips-a-vision-based-page-segmentation-algorithm/

[4] Facebook, Inc. 2013. Facebook Reports Fourth Quarter and Full Year 2012 Results. https://investor.fb.com/investor-news/press-release-details/2013/FacebookReports-Fourth-Quarter-and-Full-Year-2012-Results/default.aspx

[5] Facebook, Inc. 2021. Facebook Reports Fourth Quarter and Full Year 2020 Results. https://investor.fb.com/investor-news/press-release-details/2021/FacebookReports-Fourth-Quarter-and-Full-Year-2020-Results/default.aspx

[6] Fei Sun, Dandan Song, and Lejian Liao. 2011. DOM Based Content Extraction via Text Density. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 245–254. https: //doi.org/10.1145/2009916.2009952

[7] Jiang-Ming Yang, Rui Cai, Yida Wang, Jun Zhu, Lei Zhang, and Wei-Ying Ma. 2009. Incorporating Site-Level Knowledge to Extract Structured Data from Web Forums. In *Proceedings of the 18th International Conference on World Wide Web* (Madrid, Spain) *(WWW '09)*. Association for Computing Machinery, New York, NY, USA, 181–190. https://doi.org/10.1145/1526709.1526735

[8] Kristina Lerman, Steven N. Minton, and Craig A. Knoblock. 2003. Wrapper Maintenance: A Machine Learning Approach. *J. Artif. Int. Res.* 18, 1 (Feb. 2003), 149–181.

[9] Natalie Glance, Matthew Hurst, Kamal Nigam, Matthew Siegler, Robert Stockton, and Takashi Tomokiyo. 2005. Deriving Marketing Intelligence from Online Discussion. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (Chicago, Illinois, USA) *(KDD '05)*. Association for Computing Machinery, New York, NY, USA, 419–428. https://doi.org/10.1145/1081870.1081919

[10] Reddit. 2020. Reddit's 2020 Year in Review. https://redditblog.com/2020/12/08/ reddits-2020-year-in-review/

[11] S. Pretzsch, K. Muthmann, and A. Schill. 2012. FODEX – Towards Generic Data Extraction from Web Forums. In *2012 26th International Conference on Advanced Information Networking and Applications Workshops*. 821–826. https: //doi.org/10.1109/WAINA.2012.134

[12] Suzan Verberne, Antal van den Bosch, Sander Wubben, and Emiel Krahmer. 2017. Automatic Summarization of Domain-Specific Forum Threads: Collecting Reference Data. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 253–256. https://doi.org/10. 1145/3020165.3022127

[13] Tim Weninger, William H. Hsu, and Jiawei Han. 2010. CETR: Content Extraction via Tag Ratios. In *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, North Carolina, USA) *(WWW '10)*. Association for Computing Machinery, New York, NY, USA, 971–980. https://doi.org/10.1145/1772690. 1772789

[14] Tomas Grigalis. 2013. Towards Web-Scale Structured Web Data Extraction. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining* (Rome, Italy) *(WSDM '13)*. Association for Computing Machinery, New York, NY, USA, 753–758. https://doi.org/10.1145/2433396.2433491

[15] Yanhong Zhai and Bing Liu. 2005. Web Data Extraction Based on Partial Tree Alignment. In *Proceedings of the 14th International Conference on World Wide Web* (Chiba, Japan) *(WWW '05)*. Association for Computing Machinery, New York, NY, USA, 76–85. https://doi.org/10.1145/1060745.1060761