HAIR: Towards Developing a Global Self-Updating Peer Support Group Meeting List Using Human-Aided Information Retrieval

Sabirat Rubya
University of Minnesota
Minneapolis, MN, USA
rubya001@umn.edu

Xizi Wang
University of Michigan
Ann Arbor, MI, USA
xiziwang@umich.edu

Svetlana Yarosh
University of Minnesota
Minneapolis, MN, USA
lana@umn.edu

ABSTRACT

Alcoholics Anonymous (AA) is the largest grassroots peer support group for any health condition. While AA meeting attendance is particularly important for people who are newly sober, newcomers often have trouble finding meetings because of a lack of global up-to-date meeting list due to preference for regional autonomy in AA's organizational structure. Detection of regional webpages containing meetings and extraction of day, time, and address of meetings from those pages are essential steps in making the information available and up-to-date in a global meeting list. However, varied structure of the webpages and the meetings pose challenges in achieving the goal with traditional information retrieval methods. In this paper we propose HAIR: a semi-automated human-aided information retrieval technique and explore its potential to solve this problem. We describe future directions in developing this critical tool and discuss major implications of our work in pointing to the importance of context-specific rather than context-agnostic semi-automated information retrieval techniques by conceptualizing the proposed methods and results in a broader context.

CCS CONCEPTS

• Information systems → Information retrieval → Users and interactive retrieval • Human-centered computing → Collaborative and social computing

KEYWORDS

Alcoholics Anonymous; Peer support; Classification; Information Retrieval; Human Computation; Crowdsourcing.

ACM Reference format:

Sabirat Rubya, Xizi Wang and Svetlana Yarosh. 2019. HAIR: Towards Developing a Global Self-Updating Peer Support Group Meeting List using Human-Aided Information Retrieval. In *Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR'19). ACM, Glasgow, Scotland, UK, 10 pages.* https://doi.org/10.1145/3295750.3298933

CHIIR '19, March 10–14, 2019, Glasgow, United Kingdom

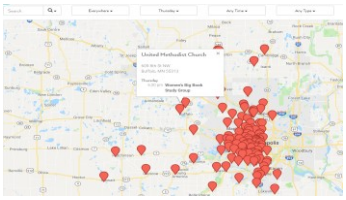
ACM ISBN 978-1-4503-6025-8/19/03...$15.00
https://doi.org/10.1145/3295750.3298933

1 INTRODUCTION

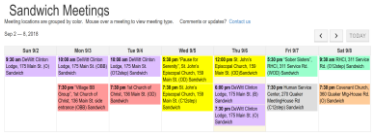
Jane just got out of a rehab. As a next step, her counselor has suggested that she attend multiple AA meetings per week in her first year of recovery. She finds a meeting finder app, installs it on her phone, and searches for a nearby meeting. She drives for 30 minutes only to find out that the meeting is not happening there anymore (turns out the meeting information in the app was gathered over a year ago)! Frustrated and confused, Jane even thinks that maybe it's a "sign" that she should go back to drinking. Luckily, she doesn't give up, after some more search, calling a hotline, and finding an updated webpage of meetings in her town she is finally able to find and attend another meeting. "How many people aren't as lucky?" she thinks.

Alcohol Use Disorder (AUD) is a prevalent and high-impact health condition. It is estimated to cost the United States $249 billion per year, with 6.2% of Americans developing an AUD [50]. Treatments for recovery from AUD are most effective when clinical intervention is coupled with long-term maintenance programs [19]. Alcoholics Anonymous (AA) is one of the most effective and popular maintenance programs [12]. It is a free peer-based social support program that offers over fifty thousand face-to-face group meetings worldwide for people struggling with AUD. While AA meetings have the potential for positive impact on people's chances of long-term recovery [31], the illustrative scenario above depicts a critical problem that could be encountered by anyone new to recovery or new to a geographic location. There is evidence that approximately 90% of alcoholics are likely to experience at least one relapse over the 4-year period following treatment [3], and Jane could have easily been one of them if she had given after being misdirected with outdated meeting data.

Individual AA groups value autonomy, and they typically operate at a local level (e.g. county, city, town, etc.). One consequence of this level of autonomy is inconsistency in how local groups make their meeting information available (e.g., different regional AA websites contain meeting information in different formats and structures) and a lack of a national or global meeting list [36,48]. Designing a static meeting finder will not solve the problem since it would soon become outdated whenever any of the regional websites updates any meeting details. Our long-term goal is to create and maintain a searchable global "self-



(a) Meetings represented as markers on a map. Clicking on a marker provides detailed information about the meeting



(b) Meeting list on a calendar

(c) List of meetings on a webpage (each meeting consists of multiple rows of an HTML table)

Day	Time	Group	Location	Address	Type
Chester					
Tuesday	8:00 PM	Chester Group of Drunks-119270	St. John's Episcopal Church	12201 Richmond St.	O/D/NS/HA
Wednesday	7:00 PM	Living Life Group-664658	Chester Presb Church	3424 West Hundred Rd	O/D/B/NS/HA
Wednesday	8:00 PM	Enon Group-128118	Bermuda United Meth.Church	2025 Florence Ave.	O/D/B/NS/HA
Thursday	8:00 PM	Voices In Recovery - 697299	St. John's Episcopal Church	12201 Richmond St.	O/SM/NS/HA
Friday	8:00 PM	Chester Big Book Group-141107	Chester Pres. Church	3424 West Hundred Rd	O/BB/NS
Saturday	11:00 AM	Common Journey	Chester Pres. Church	3424 West Hundred Rd.	O/WO/D/NS
Saturday	4:00 PM	Now or Never Meeting	Mt. Pleasant Bapt Church	13721 N. Enon Church Rd.	O/D/NS

(d) Scanned image of a meeting list (uploaded as an image on one of the regional websites)

Figure 1: Different structures of meeting pages on different regional domains

updating" AA meeting list by retrieving meeting information from all the regional websites, making it available in a searchable format, and having a systematic way for the list to self-update automatically at a regular interval.

In this paper, we approach the technical challenges in extracting meeting information from all regional websites to achieve this goal. Although there has been extensive research on information retrieval from heterogeneous unstructured sources, the automated techniques may fall short in this context due to the diversity in the format and structure of the regional websites. Additionally, retrieving complete information often requires contextual knowledge. Integration of human computation (crowdsourcing) with automated IR approaches have produced better-quality results in knowledge acquisition tasks, and thus have the potential to be applied in solving the problem under consideration. In this paper we investigate the following research questions:

RQ1: What are the information retrieval challenges of developing a global self-updating AA meeting list?

RQ2: How can we combine human computation with existing IR algorithms to provide a more accurate and scalable solution to develop the list?

We answer these questions by proposing and evaluating a semi-automated **Human-Aided Information Retrieval (HAIR)** approach to information retrieval in community contexts like AA, that may empower the members of the community to work in concert with automated classification and retrieval algorithms.

2 RELATED WORK

The process of retrieving meetings from regional AA websites includes classification to identify webpages that contain meeting lists and information retrieval to extract meeting details. We incorporated human computation techniques with the automated processes of classification and information retrieval. In this section, we describe relevant research in these three fields (classification, information retrieval, and human computation), and situate the opportunities for expanding on and combining these approaches in a single solution.

2.1 Classification of Webpages

Our proposed method includes classifying a regional AA website page as a "meeting page" (a page containing information about one or more AA meeting) or a "non-meeting page" (a page that does not).

There have been extensive investigations in classifying different types of webpages using machine learning [29,43]. Most common models built to identify, categorize, and filter information from webpages use supervised learning techniques (neural networks [2], support vector machines [18,38], etc.). In service of better online search and web content management, researchers constructed features based on the text content of URL addresses [11], structure and content of webpages (e.g., [6,34]), and aspects of neighboring pages [33,41]. We primarily followed methods from the first two categories to generate features for our machine learning model, as those are more relevant in this context. For the classification task (i.e., distinguishing a webpage containing a meeting list from a page that does not) we applied Random Forest with Bagging, since research suggests that this might be the most accurate technique in this kind of imbalanced binary classification problems [1].

However, applying these techniques off-the-shelf is unlikely to provide expected accuracy in developing a self-updating global meeting list due to two reasons. The first reason is the diversity in the structure of local AA websites. There are many different examples of structures, such as regional AA domains having seven different meeting pages (one for each day of the week) that need to be classified correctly to get a complete set of meeting pages for that area. However, these pages may seem dissimilar when fed to a model for classification, since there may be a day with dozens of meetings versus another day with just one or two meetings. The second reason involves contextual challenges in identifying meeting pages. AA meetings are not the only type of information page that includes addresses and times[1] (e.g., office hours, non-recurring organizational meetings which are not

[1] Example: http://www.utahaa.org/calendar/calendar.php

open to all, etc.). Automated processes do not contain domain knowledge and may fail to classify these pages accurately.

2.2 Information Extraction from Unstructured and Semi-Structured Sources

To extract meeting information (e.g., day, time, address) from meeting pages, we build on prior work from Information Retrieval (IR), particularly pattern detection techniques. Regional pages vary drastically in the type and structure of meetings on them (see Figure 1 for examples), necessitating our work to be situated in research on IR techniques from heterogenous unstructured sources.

Previously researchers have used natural language processing (NLP) [14], pattern matching [28], and machine learning [5,13,22], etc. for extracting data from heterogeneous semi-structured or unstructured sources. Popular applications of these techniques, "named entity recognition" [30] and "temporal information extraction [26]," are relevant here, since we want to extract specific temporal entities (i.e., address, day, and time) associated with each AA meeting. Researchers have recently used natural language to extract temporal information (e.g., events) from unstructured text in social networking sites [35] or Wikipedia [20,44]. We followed and extended a modified version of the algorithm developed by Chang *et al.* [5], in which the treelike structure of webpages and repetitive patterns were utilized to extract multiple records of similar type.

The above techniques may be used as a potential solution for retrieving meeting information, however two challenges reduce the success of off-the-shelf solutions. The first challenge is in the diversity in the format structures of each "meeting page." Meetings may be represented in different structures on different pages, such as an HTML table row or combination of table rows, an event on a calendar, or even text on a scanned image. Automated algorithms may struggle to find and relate the correct details to a specific meeting. The second challenge involves the role of context. Complete retrieval often depends on knowing how the page was accessed. For instance, some meeting pages provide an option of a dynamic search; relevant information about a meeting (e.g., time, area, type, etc.) may come from a selected option of a dropdown list. Existing methods may fail to understand this context.

2.3 Combining Human & Machine Intelligence

Due to the limitations discussed in the previous subsections, we propose to combine human computation and machine intelligence to get a better result than either could achieve alone. To leverage the "wisdom of the crowd," we suggest recruiting crowd workers from a crowdsourcing marketplace to verify both the classified webpages and the automatically extracted meeting information and to identify missing information [39].

Human computation is a computer science technique in which a computational process performs its function by outsourcing certain steps to humans [46]. Recent research has been particularly interested in systems that leverage human and machine intelligence independently and/or in sequence, combining results from both sources [40,47]. For example in the field of machine learning, crowd workers have been employed to generate ground truth data and features and to validate results after classification [7,9]. Alternatively, other systems have utilized machine learning and classification techniques to filter low quality crowdsourced data [25,42,45]. However, it is not until recently that crowdsourcing has been combined with machine computations to improve accuracy of tasks that are difficult for either to perform independently [10]. This combination has been proved efficient in solving problems like video-captioning [16], cheating detection [24], acquiring street-level accessibility information [15], and building intelligent environments [21]. Researchers have also considered utilizing crowdsourcing to improve results in solving problems like data wrangling [17], collating data from multiple sources [8], and knowledge acquisition [27,47]. This process of pipelining computation results with human verification is similar to the concept of "centaur chess" where each human player reviews possible results of candidate moves generated by a machine intelligence "teammate" before making the final decision and selecting the team's move [4]. Centaur chess teams outperform both humans and computers acting alone.

The limitations of existing automated IR techniques make our problem a good fit for combining human computing with machine intelligence. Classifying webpages and retrieving information from probable meeting pages through algorithms may not always provide accurate and expected results, when such information is crucial. Moreover, incorporating human input with machine results has proven effective in cases where contextual knowledge is important. Therefore, we propose HAIR: a human-aided information retrieval technique to address information retrieval challenges in the context of developing a global self-updating peer support meeting list for AA.

3 HAIR: HUMAN-AIDED INFORMATION RETRIEVAL

In this section, we describe the process of obtaining the initial dataset (i.e., pages from regional AA websites), followed by the two steps of HAIR to develop the meeting list: classification of pages and retrieval of meeting records from meeting pages.

3.1. Obtaining the Initial Webpages

The official website of AA lists the homepage URLs for all regional AA domains by the US state names [51]. Each of these domains usually consists of pages describing the AA program, weekly meetings in that region, non-recurring events, resources for newcomers in recovery, etc. We wrote a web scraping script using the Python Selenium library that:

1. Extracts homepage URLs of the regional websites
2. Recursively traverses through the links referenced by each homepage using depth-first search up to three levels. By manual inspection, we determined that meeting pages appeared within three levels of the homepage

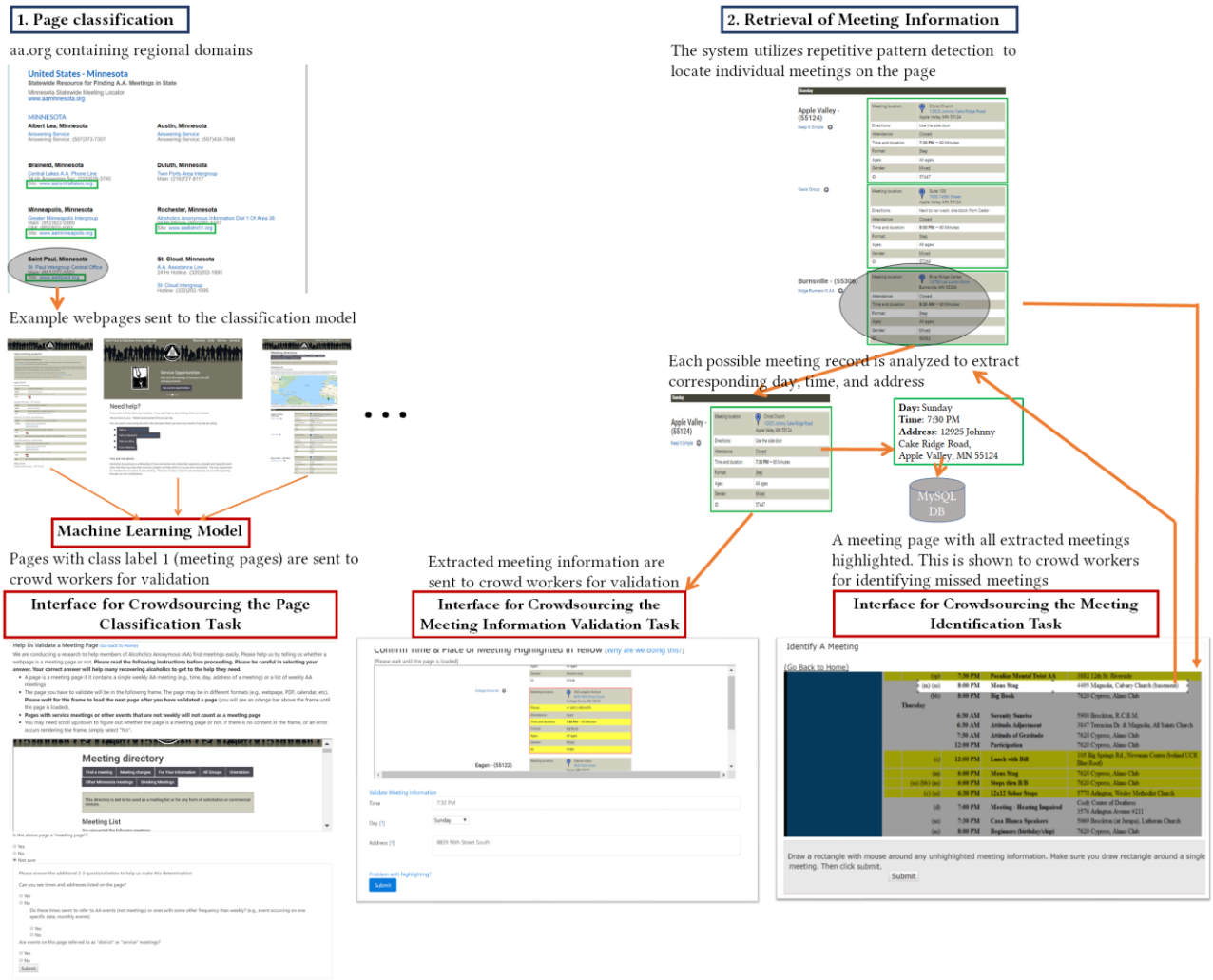


Figure 2: Pipeline of HAIR (the page classification phase is on left, and the meeting information retrieval phase is on right)

3. Determines the MIME type for each of the pages and saves the URLs along with their types in a MySQL database

Following this method, we obtained over 10,000 URLs by scraping 385 domains. This number does not match with the total number of AA regions throughout the US, since 207 regions listed their hotline numbers only. After filtering and removal of duplicates, we ended up with 9468 pages in total.

3.2 Classification of Pages

This phase consists of automatic classification of pages as "meeting page" or "non-meeting page" using a machine learning classifier and integrating human computation with the output of the classification model.

3.2.1 Classification of Pages using Machine Learning

Since AA websites contain other resources for recovery, only a small subset of the extracted URLs would refer to a meeting list. In order to efficiently reduce the number of pages to look for meeting records we developed a binary classification model.

Ground Truth and Features. We selected a random subset containing 642 pages to train the classifier, making sure to pick at least one page from each regional domain. Each page was classified manually and independently by at least two researchers, with a 0 (non-meeting page) or a 1 (meeting page) value as the class. Cases of disagreement were discussed among researchers until consensus was reached. 452 and 190 pages were classified as nonmeeting pages and meeting pages respectively.

From our experience of manually classifying hundreds of pages, we developed a list of possible page features to use in the classification. The set of final features included:

- presence of the word "meeting" in the URL (e.g., http://aaminneapolis.org/meetings)
- number of occurrences of word "meeting" in the text
- proportion of occurrences of text structured as a time
- proportion of occurrences of text structured as an address
- number of occurrences of weekday names in the text

We considered the proportions of days, times, and addressed since meeting pages often contain a larger number of times and addresses than the non-meeting pages regardless the format of the page. We considered proportions instead of raw numbers since number of meetings on a page can vary from very few to a

few hundreds. A Python script extracted these features from the 9468 pages and saved them in the database. We applied appropriate methods for scraping different formats of webpages based on the MIME type. We considered eight different formats: HTML, PDF, Google Calendar, Google Map, Document (.doc, .docx, .xls, or .xlsx), Google doc, Google sheet, and images.

Classification. This is an imbalanced classification problem [23] and we determined that the *recall* of the model is more important to consider in this context than precision. Misclassifying a meeting page may result in multiple meetings missing in the final global list. If a meeting page is misclassified with a high confidence (more on section 3.2.2), this page will not further be sent for crowdsourcing and hence this error cannot be recovered in latter stages of the pipeline. On the other hand, a misclassified non-meeting page will be either sent to crowd workers (if confidence is low) or checked for presence of meetings in the next stage and eventually removed from consideration. Therefore, we cannot overlook the cost of misclassifying the meeting pages which are a small subset of the test instances. We tried out several classification algorithms [1] using the Python WEKA wrapper to determine which one performs better with respect to recall and overall F1 score, selecting random bagging with forest.

3.2.2 Integrating Human Computation

Determining Which Pages to Send to Crowd Workers. We used a prediction margin with each classified instance, rating the confidence of the classifier associated with that prediction. We selected a cutoff value where the set of instances below the margin were sent for crowdsourcing and the remaining were accepted as correct classification. We sorted the pages by prediction margin, inspected them to determine a threshold confidence that minimizes both the size of the set below the threshold (needing crowdsourcing) and the number of errors above the threshold (number of missed meetings), and selected the cutoff to be 90%.

Interface of the Page Classification Task. We created a HIT (Human Intelligence Task) on Amazon Mechanical Turk (mTurk) [52], a crowdsourcing internet marketplace. The HIT description included instructions, examples of meeting and non-meeting pages, and a link to the web interface that we developed for the crowd workers to label meeting pages (see the interface for page classification task in Figure 2). The web interface was developed using the Python Flask framework. For each HIT, it showed 20 random pages sequentially and asked the worker if it was a meeting page with "yes", "no", and "not sure" options. If workers selected "not sure," they had to answer two or three additional questions to help us determine the label.

Recruiting Workers and Filtering Results. We conducted several pilot experiments (both with mTurkers and non-mTurkers who were undergraduate students) to refine the task design and description, to determine the pay per task, and to define a reasonable minimum time of task completion.

Based on the number of pages needing to be crowdsourced (section 4.1), we had 971 assignments of the task. Each HIT had a payment of 50 cents to fulfill the minimum state wage requirement. We followed the majority voting scheme, with five different workers labeling each page to determine the final class [10]. For ensuring label quality, we used best practices from previous work [49]. We only accepted answers from workers with greater than 50% accuracy. Additionally, we considered the amount of time (a minimum of 120 seconds based on the pilot experiments) a worker took to complete a HIT to avoid answers from untrustworthy workers. We disregarded and reassigned 51 HITs that did not meet this requirement.

3.3 Retrieval of Meeting Information

This phase consists of an automated pattern detection approach to retrieve meeting records from the pages obtained from the previous phase and crowdsourcing to identify missed meetings if any and to validate the retrieved meetings.

There is an absence of confident ground truth of meeting records for the global list, since there are about 60 thousand meetings overall. As a proof of concept, we applied the retrieval and crowdsourcing to the meeting pages from five different regional websites in Minnesota. The ground truth consisted of 1892 meetings from eight meeting pages in these domains.

3.3.1. Automated Approaches for Meeting Retrieval

We applied a pattern detection-based approach for identifying meeting location on meeting pages and a regular expression approach to extract corresponding meeting records.

Identification of Meeting Location on Meeting Pages. We used a combination of pattern detection techniques from existing Python libraries and from prior literature. This step was necessary for pinpointing individual meeting records in a list of meetings. For the webpages identified as meeting pages, we applied repetitive pattern detection technique described in [5]. The HTML content of the webpages was represented as a tree of HTML tags and similar patterns were extracted. About 60% of the meeting pages were of this type. For PDFs we used the Python PDFMiner library to determine text segments along with their coordinates. Any other types of document files were converted to PDF and the same approach was followed. We utilized the Calendar API to extract multiple events from the meeting pages where a Google Calendar was located. We analyzed the image files using the Python Tesseract library and extracted coordinates of text segments along with their content.

Extraction of Meeting Information. We checked for presence of regular expressions of time, day, or address in the repetitive patterns. If any of these aspects was missing from a pattern (e.g., the page may be accessed by selecting a particular day from a dropdown and we have to associate the day record with each of the meetings on the page), we searched that aspect in the siblings and then in the parent of that pattern recursively in the pattern tree. All events from Google Calendars were initially considered as meetings. The text segments were checked with regular expressions from PDFs and images. We considered the coordinates of the neighboring text segments for missing aspect and assigned it from the segment that is in nearest (Euclidean) distance. All records were saved in the database.

3.3.2 Integrating Human Computation

We integrated crowdsourcing to edit incorrectly fetched information by the automated approach (validation) and to retrieve

	Accuracy (%)	Recall (%)	F1 score (%)	#of meeting pages
ML	68	90.5	73.8	4268
ML+HC	80.4	94	77.8	1275

Table 1: Performance of the page classification

	Classification Result	<90% Confidence
Meeting pages	4268	3569
Non-meeting pages	4558	312

Table 2: Number of pages needing human input

meetings that were not detected (identification). We created two different HITs on mTurk and developed corresponding interfaces using the Python Flask framework.

Meeting Validation. The interface sequentially showed 10 random meetings highlighted in yellow on corresponding webpages and asked the worker if it was a meeting record, with "yes" and "no" options (Figure 2). If workers selected "no" for a meeting, they were advanced to the next record. Otherwise, they were prompted to edit the automatically extracted details of the meeting if these did not match with highlighted record.

Meeting Identification. The interface sequentially showed 20 random meeting pages with the retrieved meetings highlighted in yellow and asked the workers if they noticed any meeting as not highlighted, with "yes" and "no" options. If workers selected "no" for a page, they were advanced to the next page. Otherwise, they would be asked to select a rectangle around any un-highlighted single meeting. We then retrieved corresponding information from the answers and fed this record to the validation task. Therefore, the next worker who seeing this page would have this meeting highlighted. We continued the task until three workers had selected "no" for all the meeting pages, meaning all meetings on those pages were retrieved.

Recruiting Workers and Filtering Results. After a round of pilot experiments we established the pay per task to be 60 cents for both types of HIT, and the minimum HIT completion time to accept an answer to be 150 seconds for the meeting validation HIT and 120 seconds for the meeting identification HIT. We had 410 assignments of the meeting validation HIT and 75 assignments for the meeting identification HIT. We followed majority voting scheme where three different workers labeled each meeting. However, for edited meeting information from the validation task, we considered both majority votes and completeness (e.g., a meeting address with vs without a zip). We had to disregard and reassign answers from 20 validation HIT and 31 identification HIT that did not meet this requirement.

4 Results

We discuss the results of classification of pages and retrieval of meeting information and compare the results before and after enhancing with crowdsourcing. Additionally, we explain the results in terms of the effects in the context of AA.

4.1 Classification of Pages

The classifier reduced the number of pages to send to the crowd workers for meeting retrieval, while crowdsourcing substantially improved the accuracy of the classifier.

	Accurately Identified Meetings	Meetings identified with partially correct info (e.g., wrong time, day, or address)	Unidentified Meetings
IR	72.2 (1366)	10.9 (206)	16.9 (320)
IR+HC	94.8 (1794)	3.2 (60)	2.0 (38)

Table 3: % (number) of meetings from the Minnesota domains

We classified 8826 test pages (since 642 of the 9468 pages were used as training instances) with the supervised machine learning model (ML). The classifier output 4268 pages as "meeting pages" (with ≥90% confident about 699 pages). The accuracy of the model is not as high as recall since we tried to maximize recall. Similarly, the classifier labels pages as non-meeting with more confidence so that there is a very low probability of misclassifying a meeting page (Table 2). Nonetheless, when we looked for sources of errors we noticed that many of the misclassified meeting pages had very few meeting records and the proportions of time and address present in the text content of those pages were lower compared to other meeting pages. Moreover, non-meeting pages listing events or other details that look like meetings (e.g., office hours) were misclassified as meeting pages.

For enhancing the model with human computation (HC), we selected the 3569 meeting pages and 312 non-meeting pages for which the classifier's confidence was less than 90%. We noticed that the "not sure" option was very infrequent in the answers and we did not observe any cases where we had to consider this option. After filtering the crowdsourced answers, the number of total meeting pages was reduced to 1275. Therefore, the workers labeled 576 pages out of 3881 pages as meeting pages. The performance of the combined approach showed significantly better results in terms of recall and F1 score (see Table 1). However, we found out that the list of meeting pages still included pages with non-recurring organizational event pages, or statistics. Possible reasons for this are workers not being enough careful, or their lack of contextual knowledge.

There was a synergistic effect of combining the ML and HC. First, the automated classification reduced the number of pages needing human input from 8826 to 3881 (by about 56%), thus reduced the cost of crowdsourcing. If we reduced the confidence threshold, there would be even fewer pages, however, we may have missed more meeting pages. Second, the ML only approach would result in misclassification of more than 30% of the pages, resulting in missing a relatively high number of meetings.

Explanation in terms of the Context of Recovery

The ML only approach would have missed more than 800 meeting pages, completely making their meetings unavailable to people who need the help desperately. Although these pages usually consist of relatively small number of meetings, even pages with one meeting is important since it might be the only meeting happening in a rural area. Crowdsourcing helped identify many of these pages. On the other hand, many non-meeting pages were misclassified with ML only approach, potentially confusing, frustrating, and misdirecting people at a critical stage of their recovery. Crowdsourcing reduced this impact substantially by detecting 2993 such non-meeting pages.

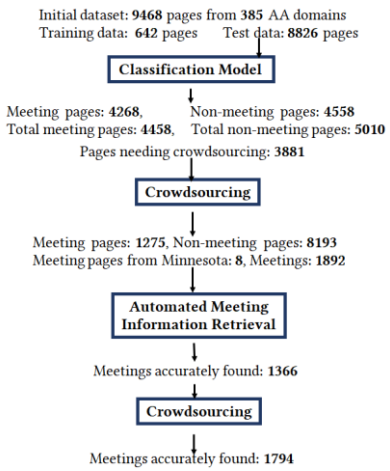


Figure 3: Summary of results

4.2 Retrieval of Meeting Information

We applied pattern-detection- and regular-expression-based approaches to identify and retrieve meeting records from pages obtained from the previous phase. Crowdsourcing complemented the automated approaches by validating information and identifying missed records.

Websites of the five regions from Minnesota had almost same variety in format of pages as the set of all pages and thus can be considered as a representative sample. Page classification phase produced 11 meeting pages from these five. However, while extracting ground truth meetings, we discarded three pages that had zero meetings. The pattern detection approach (IR) was able to detect 1572 meetings (correctly or with partially correct information) out of 1892 meeting records. However, while looking for sources of probable errors, we noticed that in the cases of attaching information present in nearby co-ordinates or in a neighboring node in the HTML tree (e.g., meetings listed by days on a page) the pattern detection often failed resulting in 16.9% of the meetings unidentified (Table-3).

Due to the variety in representation of time, day, and address and due to meetings with incomplete address information or simply listing a church as the address, regular expression did not work accurately in every case, resulting the number of correctly identified meetings to be 1365. Although only 11% of the meetings fell in the category of partially identified, it is very high if considered in the context of recovery, which we describe shortly.

Human computation (HC) substantially improved the percentage of correctly identified meetings by recognizing 282 additional meetings and reduced the percentage of meetings with wrong information by updating 146 meeting records (see the bottom row of Table 3).

There was a synergistic effect of combining the two techniques. First, the results from the combined techniques could enhance the accuracy of the page classification step. For example, there were pages where no record was marked in the IR or HC as meeting information. Clearly these are misclassified meeting pages and there is a potential for feeding the pages back to the previous phase. Second, the IR only approach would have missed relatively high number of meetings and included many meetings with potentially wrong information.

Explanation in terms of the Context of Recovery

The 3rd column of Table 3 may be considered as the most important metrics to evaluate the proposed approach in terms of effect on people in recovery. For example, the unidentified meetings could include the only meeting or a very popular meeting in a particular area and failure to find that is frustrating. However, the 2nd column refers to the fact that, after automated extraction about 11% meetings would consist of wrong information, creating confusion and potentially sending people to a location where there is no meeting. In other words, going back to the hypothetical example that we started with, this result implies that Jane has a good chance of ending up in a wrong location regardless. The results from both phases are summarized in Figure 3.

5 DISCUSSION

In this section, we conceptualize this work in a broader IR context and discuss future directions towards developing the global self-updating meeting list.

5.1 Leveraging the Specifics of Context

Prior work has investigated different IR techniques that consider human-in-the-loop (e.g., [8,27]). However, most work takes a context-agnostic approach to include human computation and make decisions about information retrieval strategies. In contrast, recovery is a high-impact context where specific decisions and approaches may lead to significant positive or negative impact on people's lives. We considered the context in:

Prioritizing errors. The classification model was selected to prioritize recall over precision. This idea is espoused in many imbalanced classification problems particularly involving fraud detection (e.g., [32]). In the context of AA, obtaining all the meeting pages was essential, even though it added extra pages to send for human input. We designed the latter phases in a way that eventually discarded those misclassified non-meeting pages.

Additionally, we considered what the percentage of unidentified meetings and the percentage of meetings identified with partially correct information would mean to someone in recovery. The latter one (while conceptually "partially correct") may have a bigger negative impact on members of AA, since it might send someone at a time or to a place where there is no meeting. That is why we took both into consideration and designed the "meeting validation" and the "meeting identification" tasks separately to ensure correctness of the extracted meetings as well as to minimize the number of unidentified meetings.

Designing for eventual user value. Our long-term goal is to provide substantial value to AA members. While in a less critical context, it would be reasonable to simply identify and catalog local pages, that may not provide sufficient value for newcomers who may already be experiencing information overload [37]. Our motivation for including the meeting retrieval phase is to produce a complete list of searchable meetings and to present them in a navigational interface that provides information in a more scalable, geographically, and temporally relevant way.

Overall, this points to the idea of moving away from context-agnostic IR methods towards ones bespoke to work better in this specific high-impact context. Building on this design insight, we plan to involve people in recovery instead of mTurk workers for human computation. One significant source of error in page classification was both the classifier and the mTurkers failing to distinguish organizational (area) meetings and the actual weekly AA meetings. Since people in recovery are more familiar with the distinction between these two, involving them will potentially improve the performance of the crowdsourcing step.

5.2 Assumptions about Ground Truth

Currently, we assume that the meeting information on any regional website is accurate, while reality may not reflect this. One possible future direction to solve this problem is to extend the human-in-the loop part of the HAIR into the physical world. Although this paper does not focus on the sustainability of the developed meeting list (more in next section), we hope to develop a meeting finder app from the extracted meetings and make it public and available for use by people in recovery. We will then have opportunities to send the app users on physical world "mission" in their local area to identify and modify possible inaccuracies on regional pages. For instance, while periodically extracting meeting information, we may note that a page has not been updated in over six months, pointing to the possibility of the page being outdated. The app may ask users searching for meetings in that area to volunteer to validate information by going to a listed meeting and confirming if that is still existent. The idea of this future work implies that human-aided information retrieval can be implemented not only online but also in the real world.

5.3 Sustainability of HAIR

This work focuses on the technical challenges of developing the meeting list. However, in future, we need to consider the sustainability of the obtained list since we are not only interested in getting a snapshot of meetings at a single point in time, but also plan to enable it to sustainably self-update.

5.3.1 Future Direction in Allowing the List to "Self-update"

To make the list self-updating one approach is to automatically run the proposed methods in a regular interval and obtain the most recent version of the meetings. We plan to update the meeting information once a month by applying the following approaches that reduce the cost of crowdsourcing:

1. Instead of classifying all webpages, we will first look for new pages added to the regional websites and pages that have changed since the last update, and send only those pages to the ML model.
2. Instead of all the pages from the previous version of the list, we will now send only the set of pages for crowdsourcing that are obtained from step 1 above.
3. After automated extraction of meeting records we will match with crowd answers from the previous version and select appropriate subset to crowdsource.
4. Before crowdsourcing, we will check the day, time, address of all meetings in a region to remove possible duplicates.

Additionally, to make information more accessible, we have developed a web API that external applications or websites can send request to fetch the set of current meeting pages and the extracted meetings from the database.

5.3.2 Consequences of Self-Updating

Some errors may disappear with time when continual updates are performed (e.g., if the crowd workers identify a wrong meeting page or a wrong meeting once, the next time the same page or meeting will possibly be evaluated by a different set of workers, creating a probability of amending the error). On the contrary, if a meeting page remains unidentified and does not change frequently, we will not consider it in subsequent self-updating processes, and the error will compound.

As we transition to involve volunteers (AA members) in crowdsourcing reducing the monetary cost to zero, the projected cost must be calculated in terms of human time. If we continue to update the list in a one-month interval, considering the current average task completion time of mTurkers, the cost would be about 96 hours human-time per update. However, it will be significantly reduced when we incorporate the steps described above in subsequent updates. AA members who have many years in recovery perform service work for the community, which creates opportunities for us to offer the crowdsourcing as a service that could enhance their recovery while also providing valuable information for newcomers.

6 CONCLUSION

Due to tradition of regional autonomy in its organizational structure, AA currently lacks a global or national meeting list. This makes it difficult for newcomers in recovery or people traveling to a new area to find the support they need. With a goal of creating a navigational meeting finder app which includes all AA meetings throughout the United States, we propose HAIR: a semi-automated information retrieval approach that classifies webpages of different structures and formats and extracts meeting information from them combining machine learning, pattern detection, and crowdsourcing techniques. We conclude that crowdsourcing has potential to be applied in concert with IR techniques in this high-impact context, as it substantially improves the accuracy of the automated approaches by being able to identify many meeting pages and to modify wrong meeting information. Additionally, a major implication of our work is pointing to the importance of context-specific rather than context-agnostic semi-automated IR methods.

ACKNOWLEDGMENTS

We would like to thank Thomas Crumrine for helping us develop the machine learning model. We also thank Zachary Levonian and Loren Terveen for their useful feedback on the paper. This work was funded by the NSF grants (1464376 and 1651575).

REFERENCES

[1] Ansam A. AbdulHussien. 2017. Comparison of Machine Learning Algorithms to Classify Web Pages. *International Journal of*

Advanced Computer Science and Applications (ijacsa) 8, 11. https://doi.org/10.14569/IJACSA.2017.081127

[2] Anagnostopoulos, C. Anagnostopoulos, V. Loumos, and E. Kayafas. 2004. Classifying Web pages employing a probabilistic neural network. *IEE Proceedings - Software* 151, 3: 139–150. https://doi.org/10.1049/ip-sen:20040121

[3] Howard C. Becker. 2008. Alcohol dependence, withdrawal, and relapse. *Alcohol Research & Health* 31, 4: 348–361.

[4] Mike Cassidy. 2014. Centaur Chess Shows Power of Teaming Human and Machine. *Huffington Post*. Retrieved August 13, 2018 from https://www.huffingtonpost.com/mike-cassidy/centaur-chess-shows-power_b_6383606.html

[5] Chia-Hui Chang and Shao-Chen Lui. 2001. IEPAD: Information Extraction Based on Pattern Discovery. In *Proceedings of the 10th International Conference on World Wide Web* (WWW '01), 681–688. https://doi.org/10.1145/371920.372182

[6] Michael Chau and Hsinchun Chen. 2008. A Machine Learning Approach to Web Page Filtering Using Content and Structure Analysis. *Decis. Support Syst.* 44, 2: 482–494. https://doi.org/10.1016/j.dss.2007.06.002

[7] Yinlin Chen, Paul Logasa Bogen II, Haowei Hsieh, Edward A. Fox, and Lillian N. Cassel. 2012. Categorization of Computing Education Resources with Utilization of Crowdsourcing. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '12), 121–124. https://doi.org/10.1145/2232817.2232840

[8] Zhe Chen. 2014. A Semiautomatic Approach for Accurate and Low-Effort Spreadsheet Data Extraction.

[9] Justin Cheng and Michael S. Bernstein. 2015. Flock: Hybrid Crowd-Machine Learning Classifiers. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '15), 600–611. https://doi.org/10.1145/2675133.2675214

[10] Ed H. Chi. 2017. Technical Perspective: Humans and Computers Working Together on Hard Tasks. *Commun. ACM* 60, 9: 92–92. https://doi.org/10.1145/3068614

[11] M. I. Devi, R. Rajaram, and K. Selvakuberan. 2007. Machine Learning Techniques for Automated Web Page Classification Using URL Features. In *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, 116–120. https://doi.org/10.1109/ICCIMA.2007.342

[12] M. Ferri, L. Amato, and M. Davoli. 2006. Alcoholics Anonymous and other 12-step programmes for alcohol dependence. *The Cochrane Database of Systematic Reviews*, 3: CD005032. https://doi.org/10.1002/14651858.CD005032.pub2

[13] Dayne Freitag. 2000. Machine Learning for Information Extraction in Informal Domains. *Machine Learning* 39, 2–3: 169–202. https://doi.org/10.1023/A:1007601113994

[14] F. S. Gharehchopogh and Z. A. Khalifelu. 2011. Analysis and evaluation of unstructured data: text mining versus natural language processing. In *2011 5th International Conference on Application of Information and Communication Technologies (AICT)*, 1–4. https://doi.org/10.1109/ICAICT.2011.6111017

[15] Kotaro Hara and Jon E. Froehlich. 2015. Characterizing and Visualizing Physical World Accessibility at Scale Using Crowdsourcing, Computer Vision, and Machine Learning. *SIGACCESS Access. Comput.*, 113: 13–21. https://doi.org/10.1145/2850440.2850442

[16] Yun Huang, Yifeng Huang, Na Xue, and Jeffrey P. Bigham. 2017. Leveraging Complementary Contributions of Different Workers for Efficient Crowdsourcing of Video Captions. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17), 4617–4626. https://doi.org/10.1145/3025453.3026032

[17] Zhongjun Jin, Michael R. Anderson, Michael Cafarella, and H. V. Jagadish. 2017. Foofah: Transforming Data By Example. In *Proceedings of the 2017 ACM International Conference on Management of Data* (SIGMOD '17), 683–698. https://doi.org/10.1145/3035918.3064034

[18] Thorsten Joachims. 1998. Text categorization with Support Vector Machines: Learning with many relevant features. In *Machine Learning: ECML-98*, Claire Nédellec and Céline Rouveirol (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 137–142. https://doi.org/10.1007/BFb0026683

[19] Herbert D. Kleber, Roger D. Weiss, Raymond F. Anton, Tony P. George, Shelly F. Greenfield, Thomas R. Kosten, Charles P. O'Brien, Bruce J. Rounsaville, Eric C. Strain, Douglas M. Ziedonis, Grace Hennessy, Hilary Smith Connery, John S. McIntyre, Sara C. Charles, Daniel J. Anzia, Ian A. Cook, Molly T. Finnerty, Bradley R. Johnson, James E. Nininger, Paul Summergrad, Sherwyn M. Woods, Joel Yager, Robert Pyles, C. Deborah Cross, Roger Peele, John P. D. Shemo, Lawrence Lurie, R. Dale Walker, Mary Ann Barnovitz, Sheila Hafter Gray, Sunil Saxena, Tina Tonnu, Robert Kunkle, Amy B. Albert, Laura J. Fochtmann, Claudia Hart, Darrel Regier, Work Group on Substance Use Disorders, American Psychiatric Association, and Steering Committee on Practice Guidelines. 2007. Treatment of patients with substance use disorders, second edition. American Psychiatric Association. *The American Journal of Psychiatry* 164, 4 Suppl: 5–123.

[20] Erdal Kuzey and Gerhard Weikum. 2012. Extraction of Temporal Facts and Events from Wikipedia. In *Proceedings of the 2Nd Temporal Web Analytics Workshop* (TempWeb '12), 25–32. https://doi.org/10.1145/2169095.2169101

[21] Gierad Laput, Walter S. Lasecki, Jason Wiese, Robert Xiao, Jeffrey P. Bigham, and Chris Harrison. 2015. Zensors: Adaptive, Rapidly Deployable, Human-Intelligent Sensor Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 1935–1944. https://doi.org/10.1145/2702123.2702416

[22] Wendy Lehnert, Stephen Soderland, David Aronow, FangFang Feng, and Avinoam Shmueli. 1995. Inductive text classification for medical applications. *Journal of Experimental & Theoretical Artificial Intelligence* 7, 1: 49–80. https://doi.org/10.1080/09528139508953800

[23] Camelia Lemnaru and Rodica Potolea. 2012. Imbalanced Classification Problems: Systematic Study, Issues and Best Practices. In *Enterprise Information Systems* (Lecture Notes in Business Information Processing), 35–50.

[24] Xuanchong Li, Kai-min Chang, Yueran Yuan, and Alexander Hauptmann. 2015. Massive Open Online Proctor: Protecting the Credibility of MOOCs Certificates. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '15), 1129–1137. https://doi.org/10.1145/2675133.2675245

[25] Henry Lieberman, Karthik Dinakar, and Birago Jones. 2013. Crowdsourced Ethics with Personalized Story Matching. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '13), 709–714. https://doi.org/10.1145/2468356.2468481

[26] Xiao Ling and Daniel S. Weld. 2010. Temporal Information Extraction. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*. Retrieved September 21, 2018 from https://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/view/1805

[27] Huaxi Liu, Ning Wang, and Xiangran Ren. 2015. CrowdSR: A Crowd Enabled System for Semantic Recovering of Web Tables. In *Web-Age Information Management* (Lecture Notes in Computer Science), 581–583.

[28] Brian McLernon and Nicholas Kushmerick. 2006. *Transductive Pattern Learning for Information Extraction.* UNIVERSITY COLL DUBLIN (IRELAND), UNIVERSITY COLL DUBLIN (IRELAND). Retrieved August 19, 2018 from http://www.dtic.mil/docs/citations/ADA456766

[29] Dunja Mladenic. 1998. *Turning Yahoo into an Automatic Web-Page Classifier.*

[30] Raymond J. Mooney and Razvan Bunescu. 2005. Mining Knowledge from Text Using Information Extraction. *SIGKDD Explor. Newsl.* 7, 1: 3–10. https://doi.org/10.1145/1089815.1089817

[31] Rudolf H. Moos and Bernice S. Moos. 2006. Participation in Treatment and Alcoholics Anonymous: A 16-Year Follow-Up of Initially Untreated Individuals. *Journal of clinical psychology* 62, 6: 735–750. https://doi.org/10.1002/jclp.20259

[32] Clifton Phua, Damminda Alahakoon, and Vincent Lee. 2004. Minority Report in Fraud Detection: Classification of Skewed Data. *SIGKDD Explor. Newsl.* 6, 1: 50–59. https://doi.org/10.1145/1007730.1007738

[33] Xiaoguang Qi and Brian D. Davison. 2006. Knowing a Web Page by the Company It Keeps. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management* (CIKM '06), 228–237. https://doi.org/10.1145/1183614.1183650

[34] Xiaoguang Qi and Brian D. Davison. 2009. Web Page Classification: Features and Algorithms. *ACM Comput. Surv.* 41, 2: 12:1–12:31. https://doi.org/10.1145/1459352.1459357

[35] Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. 2012. Open Domain Event Extraction from Twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '12), 1104–1112. https://doi.org/10.1145/2339530.2339704

[36] Sabirat Rubya. 2017. Facilitating Peer Support for Recovery from Substance Use Disorders. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '17), 172–177. https://doi.org/10.1145/3027063.3048431

[37] Sabirat Rubya and Svetlana Yarosh. 2017. Video-Mediated Peer Support in an Online Community for Recovery from Substance Use Disorders. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17), 1454–1469. https://doi.org/10.1145/2998181.2998246

[38] E. Saraç and S. A. Özel. 2013. Web page classification using firefly optimization. In *2013 IEEE INISTA*, 1–5. https://doi.org/10.1109/INISTA.2013.6577619

[39] Thimo Schulze, Simone Krug, and Martin Schader. 2012. Workers' Task Choice in Crowdsourcing and Human Computation Markets. *ICIS 2012 Proceedings*. Retrieved from http://aisel.aisnet.org/icis2012/proceedings/ResearchInProgress/40

[40] Vinay Shashidhar, Nishant Pandey, and Varun Aggarwal. 2015. Spoken English Grading: Machine Learning with Crowd Intelligence. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '15), 2089–2097. https://doi.org/10.1145/2783258.2788595

[41] Wongkot Sriurai, Phayung Meesad, and Choochart Haruechaiyasak. 2010. Hierarchical Web Page Classification Based on a Topic Model and Neighboring Pages Integration. *arXiv:1003.1510 [cs]*. Retrieved August 13, 2018 from http://arxiv.org/abs/1003.1510

[42] Chong Sun, Narasimhan Rampalli, Frank Yang, and AnHai Doan. 2014. Chimera: Large-scale Classification Using Machine Learning, Rules, and Crowdsourcing. *Proc. VLDB Endow.* 7, 13: 1529–1540. https://doi.org/10.14778/2733004.2733024

[43] Makoto Tsukada, Takashi Washio, and Hiroshi Motoda. 2001. Automatic Web-Page Classification by Using Machine Learning Methods. In *Proceedings of the First Asia-Pacific Conference on Web Intelligence: Research and Development* (WI '01), 303–313. Retrieved January 11, 2017 from http://dl.acm.org/citation.cfm?id=645960.673927

[44] Yafang Wang, Mingjie Zhu, Lizhen Qu, Marc Spaniol, and Gerhard Weikum. 2010. Timely YAGO: Harvesting, Querying, and Visualizing Temporal Knowledge from Wikipedia. In *Proceedings of the 13th International Conference on Extending Database Technology* (EDBT '10), 697–700. https://doi.org/10.1145/1739041.1739130

[45] 4Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A. Smith, and Frederick Liu. 2016. Crowdsourcing Annotations for Websites' Privacy Policies: Can It Really Work? In *Proceedings of the 25th International Conference on World Wide Web* (WWW '16), 133–143. https://doi.org/10.1145/2872427.2883035

[46] Yang Yang, Bin B. Zhu, Rui Guo, Linjun Yang, Shipeng Li, and Nenghai Yu. 2008. A Comprehensive Human Computation Framework: With Application to Image Labeling. In *Proceedings of the 16th ACM International Conference on Multimedia* (MM '08), 479–488. https://doi.org/10.1145/1459359.1459423

[47] Dahai Yao, Hailong Sun, and Xudong Liu. 2015. Combining Crowd Contributions with Machine Learning to Detect Malicious Mobile Apps. In *Proceedings of the 7th Asia-Pacific Symposium on Internetware* (Internetware '15), 120–123. https://doi.org/10.1145/2875913.2875941

[48] Svetlana Yarosh. 2013. Shifting Dynamics or Breaking Sacred Traditions?: The Role of Technology in Twelve-step Fellowships. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), 3413–3422. https://doi.org/10.1145/2470654.2466468

[49] D. Yue, G. Yu, D. Shen, and X. Yu. 2014. A Weighted Aggregation Rule in Crowdsourcing Systems for High Result Accuracy. In *2014 IEEE 12th International Conference on Dependable, Autonomic and Secure Computing*, 265–270. https://doi.org/10.1109/DASC.2014.54

[50] Alcohol Facts and Statistics | National Institute on Alcohol Abuse and Alcoholism (NIAAA). Retrieved September 6, 2018 from https://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/alcohol-facts-and-statistics

[51] Alcoholics Anonymous: A.A. Near You. Retrieved September 23, 2018 from https://www.aa.org/pages/en_US/find-aa-resources

[52] Amazon Mechanical Turk. Retrieved September 23, 2018 from https://www.mturk.com/