Structured Labeling to Facilitate Concept Evolution in Machine Learning

Todd Kulesza[1,2], Saleema Amershi[2], Rich Caruana[2], Danyel Fisher[2], Denis Charles[2]

[1]Oregon State University
Corvallis, OR
kuleszto@eecs.oregonstate.edu

[2]Microsoft Research
Redmond, WA
{samershi, rcaruana, danyelf, cdx}@microsoft.com

ABSTRACT

Labeling data is a seemingly simple task required for training many machine learning systems, but is actually fraught with problems. This paper introduces the notion of *concept evolution*, the changing nature of a person's underlying *concept* (the abstract notion of the target class a person is labeling for, e.g., *spam* email, *travel related* web pages) which can result in inconsistent labels and thus be detrimental to machine learning. We introduce two *structured labeling* solutions, a novel technique we propose for helping people define and refine their concept in a consistent manner as they label. Through a series of five experiments, including a controlled lab study, we illustrate the impact and dynamics of concept evolution in practice and show that structured labeling helps people label more consistently in the presence of concept evolution than traditional labeling.

Author Keywords
Concept evolution; interactive machine learning

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI): User interfaces.

INTRODUCTION

Data is fundamental in machine learning. In supervised learning, a machine is trained from example data that is labeled according to some target *concept*. The result is a learned function that can predict the labels of new, unseen data. The performance of machine learning depends on the quality of the labeled data used for training. For example, spam filters are often machine-learned functions that are trained from a large corpus of emails or web pages labeled as *spam* or *not spam*. Poorly performing spam filters may admit unwanted spam or, worse yet, incorrectly classify important email or web pages as spam.

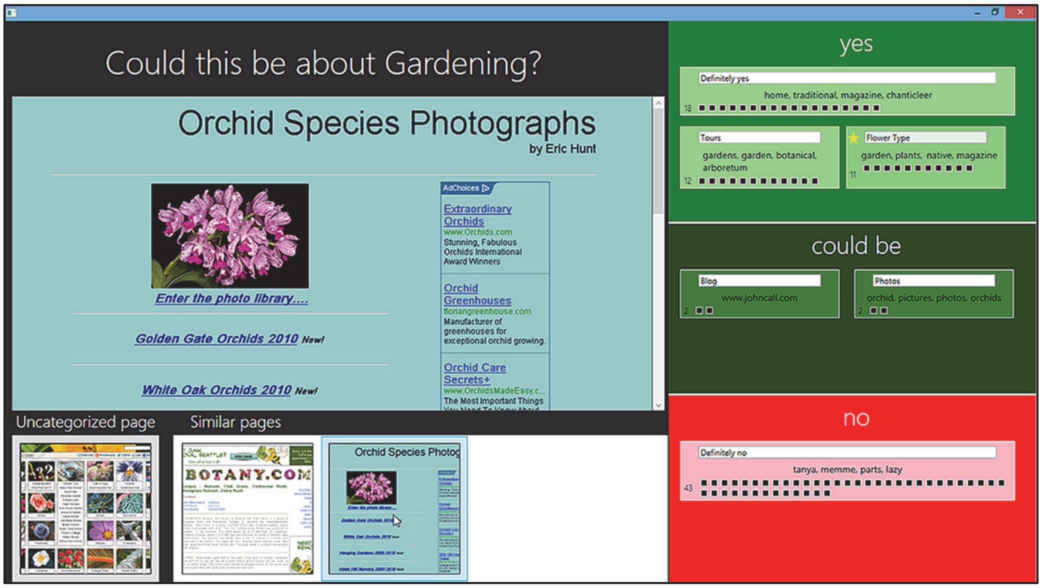
Large corporations often recruit people to label the large amounts of data machine learners need to support automated services such as ranking web search results (e.g., [16, 20]), providing recommendations (e.g., [31]), or displaying relevant ads (e.g., [34]). Additionally, interactive machine learning systems allow individual end-users to label data to improve personalized services such as email filtering and prioritization (e.g., [12, 13]), and music or movie recommendations (e.g., [35]).

While labeling data is a seemingly simple task, it is actually fraught with problems (e.g., [9, 19, 26]). Labels reflect a labeler's mapping between the data and their underlying *concept* (i.e., their abstract notion of the target class). Thus, label quality is affected by factors such as the labeler's expertise or familiarity with the concept or data, their judgment ability and attentiveness during labeling, and the ambiguity and changing distribution of the data itself.

This paper addresses a distinct problem in labeling data that we refer to as *concept evolution*. Concept evolution refers to the labeler's process of defining and refining a concept in their minds, and can result in different labels being applied to similar items due to changes in the labeler's notion of the underlying concept. In a formative study presented later in this paper, we found that people labeling a set of web pages twice with a four-week gap between labeling sessions were, on average, only 81% consistent with their initial labels. This inconsistency in labeling similar items can be harmful to machine learning, which is fundamentally based on the idea that similar inputs should have similar outputs [18]. Further, while label quality is always important in machine learning, quality is especially critical in situations where data quantity is limited (e.g., when labels are expensive to obtain or when individuals are labeling data for their own purposes, as in many interactive machine learning systems) [1].

To address the concept evolution problem, we introduce *structured labeling* (Figure 1), a novel interaction technique for helping people define and refine their concepts as they label data. Structured labeling allows people to organize their concept definition by grouping and tagging data (as much or as little as they choose) within a *traditional labeling* scheme (e.g., labeling into mutually exclusive categories such as 'yes', 'no', and 'could be'). This organization capability helps to increase label consistency by helping people

CHI 2014, April 26 - May 01 2014, Toronto, ON, Canada
Copyright 2014 ACM 978-1-4503-2473-1/14/04 $15.00.
http://dx.doi.org/10.1145/2556288.2557238



Could this be about Gardening?

Orchid Species Photographs
by Eric Hunt

Enter the photo library....

Golden Gate Orchids 2010 New!

White Oak Orchids 2010 New!

AdChoices

Extraordinary Orchids
www.Orchids.com
Stunning, Fabulous Orchids International Award Winners

Orchid Greenhouses
floriangreenhouse.com
Manufacturer of greenhouses for exceptional orchid growing.

Orchid Care Secrets+
www.OrchidsMadeEasy.c...
The Most Important Things

Uncategorized page

Similar pages

yes

Definitely yes
home, traditional, magazine, chanticleer
18

Tours
gardens, garden, botanical, arboretum
12

Flower Type
garden, plants, native, magazine
11

could be

Blog
www.johncall.com
2

Photos
orchid, pictures, photos, orchids
2

no

Definitely no
tanya, memme, parts, lazy
43

Figure 1. Our structured labeling approach allows people to group data in whatever way makes sense to them. By seeing the resulting structure, people can gain a deeper understanding of the concept they are modeling. Here, the user sees an uncategorized page (top left) and can drag it to an existing group (right), or create a new group for it. The thumbnails (bottom left) show similar pages in the dataset to help the user gauge whether creating a new group is warranted.

explicitly surface and recall labeling decisions. Further, because the structure is malleable (users can create, delete, split, and merge groups), it is well-suited for situations where users are likely to frequently refine their concept definition as they observe new data. We also present an *assisted structured labeling* version of our tool that uses visual aids and label recommendations to further assist people labeling data while their concept evolves.

This paper makes the following contributions:

- We introduce the concept evolution problem and present findings from three formative studies illustrating the impact and dynamics of concept evolution in situations where people interact with machine learning.
- We introduce two tools for *structured labeling*, a novel interaction technique for helping people evolve their concepts during labeling.
- We present a controlled experiment comparing our structured labeling tools to traditional labeling in machine learning. Our findings show that structured labeling was preferred by participants and helped them label data more consistently, but at a cost of speed.
- We present findings from a small follow up experiment showing that structured labeling helps people arrive at more consistent structuring decisions than traditional labeling when labeling the same data ten days apart.

BACKGROUND AND RELATED WORK

In this section we describe how concept evolution differs from other labeling problems in machine learning and why existing solutions to these problems therefore do not address concept evolution. We then describe related work in sensemaking and information management that inspired our structured labeling approach to concept evolution and explain how our work extends research in this area to the problem of labeling in machine learning.

Labeling Challenges and Solutions in Machine Learning

Supervised machine learning requires labeled data for training a machine learner [18]. However, there are many well-studied challenges that arise from obtaining human-labeled data. For example, labels can be *noisy*, meaning some data may be mislabeled or labels may be applied inconsistently. As a result, the machine learning community has developed noise-tolerant algorithms (e.g., [10, 30) and techniques for automatically identifying and eliminating or correcting mislabeled data (e.g., [9]). While algorithmic solutions can reduce the impact of label noise on the quality of the resulting machine learners, they do not help *users* refine a target concept in their own minds.

More recently, researchers have started to explore novel interfaces to reduce label noise. For example, Carterette et al. [11] showed that pairwise comparisons (e.g., is document *A* better than document *B*) produced more reliable relevance labels than absolute judgments. Set-wise judgments have also been explored for obtaining relevance labels (e.g., [4]). While comparison-based judgments have been shown to be easier to make than absolute judgments, relevance judgments may still evolve as people observe data [7]. Thus, techniques for soliciting labels via comparisons can still benefit from support for structuring and revisiting labeling decisions.

Another common approach to dealing with label noise is to use multiple labelers and majority voting or weighting schemes to make final label judgments (e.g., [19, 26]). Again, while techniques involving multiple labelers might help reduce label noise, they do not solve the concept evolution problem. In contrast, our structured labeling approach may benefit these multiple labeler solutions by enabling people to share their labeling decisions and potentially even converge on a target concept definition.

Furthermore, a growing class of *interactive* machine learning systems rely on labeled data from individual users [1] and therefore cannot benefit from multiple labeler solutions. Because the amount of data individuals may be willing or able to label can be much less than in multiple labeler solutions, even a few mislabeled items can be a substantial portion of the data and be detrimental to learning [8].

An even more insidious problem in data labeling is *concept drift*, where the underlying data is fundamentally changing over time [29]. An example of concept drift is a news recommender that attempts to recommend the most interesting recent news. Here, the concept of *interesting* may remain the same over time, but the data (in this case the news) is constantly drifting as a result of changing current events. Most solutions to concept drift model concepts temporally, such as by discarding or weighting information according to a moving window over the data (e.g., [27, 33) or by automatically identifying new types of data (e.g., [5, 15]). Critically, none of these solutions are intended to help a *user* refine their own idea of a concept, a problem which may be exacerbated in the presence of concept drift.

Tools for Sensemaking and Data Management

Our proposed structured labeling solution to the concept evolution problem is inspired by work in sensemaking [23], the iterative process of organizing and understanding large amounts of data. Our work is particularly related to sensemaking research for information and document management (e.g., [3, 14, 21, 24]). As with structured labeling, these tools often leverage spatial memory and visual representations to help people organize information [32]. For example, Data Mountain facilitates sensemaking and information management by enabling users to arrange documents in a 3D virtual environment [24]. Teevan et al. [28] explored several visual representations of information to aid people in finding and re-finding information. Others have explored techniques for visualizing groups of documents, such as fanning out or stacking document thumbnails and displaying textual summaries (e.g., [3, 24]).

Our assisted structured labeling tool, which employs automated visual cues and recommendations, is closely related to recent work on tools for semi-automated support of sensemaking and information management (e.g., [2, 3]. For example, iCluster [3] helps people sort documents into spatially organized groups or piles by making group recommendations for incoming documents via highlighting. Similarly, CueT [2] helps people triage network alarms (described by short snippets of information), into existing alarm groups by making group recommendations via a ranked list and accompanying confidence visualization.

All of these tools support sensemaking to facilitate personal or collaborative information consumption and management such as browsing and navigating, searching and re-finding, and sharing or distributing information. In contrast, our work on structured labeling extends sensemaking to the domain of document labeling for machine learning and demonstrates the impact of such supports on the quality of human provided labels. In addition, as our studies reveal, sensemaking for the labeling task presents unique information management problems that require novel solutions such as helping users determine if and how to organize individual documents and how to make labeling decisions.

CONCEPT EVOLUTION IN PRACTICE

To better understand concept evolution and inform the design of our proposed solution, we conducted a series of formative studies investigating concept evolution in practice (i.e., in situations involving people labeling data for machine learning systems). Observations and feedback from these studies informed our final prototypes, as discussed in the Structured Labeling and Assistance section.

Concept Evolution during Interactive Machine Learning

Even experienced machine learning practitioners evolve their concepts while labeling. We asked 11 machine learning experts from a large research organization to train binary web page classifiers via an interactive machine learning system. Each expert labeled data according to a concept of their choice, selected from a list of Open Directory Project[1] topics (e.g., *photography*, *employment*, and *math*). From a questionnaire distributed after the session, we found that nine of the participants "defined/refined their concept while interacting" with the tool. This concept evolution could be the result of viewing additional data (only three people disagreed that their concept evolved "as a result of seeing web pages") or of using other features offered by the tool (e.g., viewing errors or experimenting with different feature combinations). This suggests that multiple factors may trigger concept evolution. Interestingly, seven participants also stated "I had a clear idea about the concept I was modeling before starting". However, four of these seven also agreed or were neutral about the statement "the concept I was modeling evolved as I saw web pages", suggesting that even when people are familiar with a concept, their definition of it may still evolve.

Concept Evolution during Creation of Label Guidelines

When acquiring labeled data for the purpose of training or testing machine learning, researchers and practitioners often create guidelines for labelers in order to obtain consistent labels (e.g., [16, 20]), where *consistent* is defined as similar items having the same label. We interviewed two practitioners from a large organization with extensive experience creating such guidelines for human labelers. Both

[1] http://www.dmoz.org/

practitioners described the typical guideline creation process as iterative and evolving as a result of observing new data.

According to our interviews, a team of people would first look through example data potentially matching their target concept. Next, the team would discuss how different examples should be labeled and come up with rules to explain their decision making. These rules often targeted difficult cases (e.g., examples with multiple interpretations), explicitly stating how such cases should be labeled and giving concrete examples. Often there would be several rounds of this process until a relatively comprehensive set of rules were generated and could be passed off to labelers.

Impact of Concept Evolution on Label Consistency and Initial Feedback on Structured Labeling

We conducted a preliminary study using an early prototype of our structured labeling tool to examine the impact of concept evolution and obtain feedback about how the tool could be improved. This prototype displayed one web page at a time and asked participants to categorize the page into one of three high-level categories: 'yes', 'no', or 'could be' (i.e., this is, is not, or could be an example of concept *X*). In addition, participants could create groups within the 'could be' category and tag them to remind themselves of what they had put in the group. (Throughout this paper, *category* refers to the high-level labels 'yes', 'no', and 'could be', while *group* refers to a user-created collection of items within a category. A *tag* is a user-supplied group description.)

We asked nine of the participants from our first study to label ~200 web pages using our prototype and according to the same concept they chose for the previous study. For each participant, 75% of the pages were the same pages they had labeled during the first study (displayed in a different order). This study occurred about four weeks after the first study.

Examining the consistency of participants' high-level labels (i.e., the 'yes', 'no' and 'could be' categories) from the first study to the second revealed that, on average, participants were only 81.7% (SD=6.8%) consistent with their previous labels (Figure 2). McNemar-Bowker tests of symmetry also showed that six of the nine participants' labels changed significantly (χ^2 (3,N=156)=9.51-30.37, p<.05) from the first study. This lack of consistency means models learned from these labels will be different, even for the *same person's* definition of the *same concept*—their concept definition significantly evolved between the two labeling sessions.

In addition to finding evidence of concept evolution, this study revealed some of the benefits of structured labeling. First, all nine participants created groups (median of six groups per participant with three pages per group). We manually annotated each of the groups and determined that most groups (~76%) were topical in nature (e.g., 'computing related' or 'mathematicians' groups for the *math* concept). The rest pertained to items people wanted to revisit for different reasons, essentially deferring decisions until they had a better understanding of the data (e.g., 'mixed content', 'more info needed'). This visible organization proved popular—in one participant's own words:

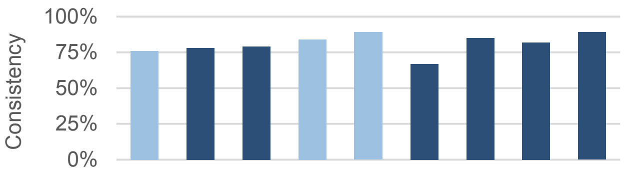


Figure 2. Label consistency (per participant) on the same data labeled approximately four weeks apart. Dark bars indicate significant differences.

"[Categorization] allows me to organize my thoughts."

Further, participants felt that seeing the structure made labeling less stressful because they could easily see and revise their labels as needed:

"I like the structuring. It's like a softer way of labeling."

STRUCTURED LABELING AND ASSISTANCE

In this section, we describe our structured labeling and assisted structured labeling prototypes and relate our design decisions back to findings from our formative studies.

Structured Labeling Prototype

Our structured labeling prototype (Figure 1) allows users to organize data within a traditional labeling scheme (e.g., mutually exclusive categories such as 'yes', 'no', and 'could be') via grouping and tagging. The system presents users with one page at a time, which they can drag to the labeling area (right in Figure 1) to create a new group or to add it an existing group. Users can manually add tags describing each group to aid in recall (e.g., 'Tours' in Figure 3).

In our formative study with an early version of our structured labeling prototype, we only allowed users to structure within the 'could be' category—we believed users would only want to structure ambiguous items. However, several of our participants requested structuring within the 'yes' and 'no' categories as well, citing a desire to preserve and revisit groups so they could later decide whether each group was part of their concept or not (i.e., moving groups between the 'could be' and 'yes' or 'no' categories):

"It's nice not to lose information in a big pile [of 'yes' or 'no' pages]."

Another reason users desired structuring within the 'yes' and 'no' categories was to make their labeling decisions visible to others:

"Comics that became movies is ambiguous, [but] you can imagine someone [else] would be interested."

Based on this feedback, we altered our prototype to support structuring in all categories. We also enabled structure editing via moving groups between categories (preserving any accompanying tags), merging groups, or moving individual items between groups. In addition, while we enabled structuring in all categories, we wanted to encourage users to focus on structuring only items they wished to revisit (and not items they felt clearly belonged to their concept or not); thus, we created default 'Definitely yes' and 'Definitely no' groups within the 'yes' and 'no' categories (Figure 1).

Assisted Structuring

During our formative study we observed participants encountering obstacles with some structuring capabilities. Here we describe additional supports we designed to help users overcome these obstacles. These supports are included as part of our *assisted structured labeling* tool.

Helping Users Recall Group Contents

Participants often had trouble remembering what they had placed in each group. While they could tag each group with a textual description, many participants did not initially make use of this feature and later regretted not taking the time to tag groups. As one participant phrased it:

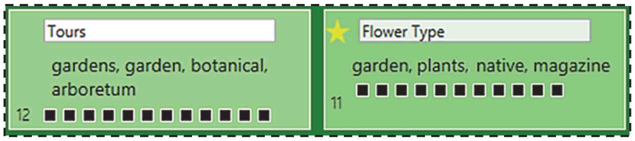
> *"Now that I want to insert [another item] I wish I had a title."*

To help users recall group contents, we augmented our structured labeling tool to automatically generate and display *textual summaries* for each group (Figure 3). Users could still manually supply tags in addition to these summaries.

We experimented with two bag-of-words approaches for creating textual summaries. Initially, we considered the content of web pages within each group as a bag-of-words (i.e., the set of words from all the pages within a group), and selected the most frequently occurring words from the bag, with *frequency* computed via the common term frequency-inverse document frequency (TF-IDF) measure [22]. However, we found that the resulting words were difficult to interpret out of context. We then turned to a corpus of search query information from a popular search engine. Thus, each web page was represented by a set of search query phrases that actual people used to find that web page via the search engine. Because such phrases are typically short and targeted, we believed they might generate clearer summaries. Therefore, we took the same approach of considering each group of web pages as a bag-of-words, this time made up of search query phrases, and selected the words with the highest TF-IDFs to display as our summaries. Intuitively, the summaries displayed the most prominent search terms used to find the web pages within each group. These summaries were updated in real-time as group contents changed.

Helping Users Decide Where to Group Items

During our formative study, we observed people having trouble deciding which group to put items in when they had several groups with related content:



Tours

gardens, garden, botanical, arboretum

12 ■■■■■■■■■■■■

Flower Type

garden, plants, native, magazine

11 ■■■■■■■■■■■

Figure 3 Our assisted structuring tool provides users with automatic summaries of each group's contents (below the user-supplied tag area) and recommends a group for the current item via an animation and yellow star indicator. The black squares indicate how many items are in each group.

> *"I remember seeing a page like this, but I can't remember which decision I made."*

To help people decide which group might be most appropriate for each new item, we added group recommendations to our structured labeling tool. Recommendations were made by computing the similarity between a new item and each group, with the most similar group recommended. We computed item-to-group similarity as the similarity between the new item and the most similar item in the group (i.e., we computed similarity between the new item and all members of a group and then selected the 'shortest-link' as the similarity value). We computed item-to-item similarity via the common cosine similarity metric over a TF-IDF representation of the contents of each item.

Group recommendations were shown in the interface using a 'wiggle' animation on the group to draw the user's attention and a static indicator visible within the recommended group (the 'star' icon in Figure 3).

Helping Users Determine When and How to Make Decisions

In our formative study, we observed that participants did not want to expend effort labeling or grouping 'outliers':

> *"If there are more than one or a few [pages] with that same property then I'll think about it, otherwise I won't."*

Other participants said that seeing multiple related items helped them decide how items should be labeled:

> *"Once you see a lot in a group, it helps you decide"*

To help people determine if an item is an outlier or one of many similar items, we included a display of the most similar unlabeled pages to the item currently being labeled (the small thumbnails displayed horizontally at the bottom of Figure 1). Similar items were identified using the same item-to-item similarity measure used to make group recommendations.

EVALUATING STRUCTURED LABELING

Our structured labeling tools are intended to help people consistently define and refine their concepts as they observe data. Therefore, we designed a controlled experiment to compare structured labeling to traditional labeling in machine learning in terms of label quality, speed, and use and preference for structuring to help define concepts.

Conditions and Tasks

Our experiment tested three interface conditions: a **manual structuring** condition with support for structured labeling (but without any automated assistance), an **assisted structuring** condition with structuring support plus automated assistance, and a **baseline** condition representing traditional labeling into mutually exclusive 'yes', 'no', and 'could be' categories and with no structuring support (i.e., the manual interface without the ability to create groups).

In order to compare subjective preferences for these three labeling techniques, we intended for every participant to use each interface. Consequently, we needed to develop three comparable—but different—labeling tasks. This is non-trivial due to the variety of factors affecting labeling and structuring decisions (e.g., concept familiarity, the inherent structure present in the data). Therefore, we prioritized the following requirements affecting our primary objectives:

- *All participants should be reasonably familiar with the target concept for each task.* This was to reduce any frustrations and delays that might be caused by lack of familiarity with a concept during labeling (e.g., a person unfamiliar with 'equestrianism' might become frustrated trying to label items as related to equestrianism or not).
- *Each task should contain the same number of items to label and roughly the same proportion of items likely belonging, not belonging, and possibly belonging to the target concept* (i.e., items likely to be labeled 'yes', 'no', and 'could be', respectively). This was to reduce any effects of label class on labeling speed, as our formative studies showed that clearly positive and negative items were typically easier and faster to label than ambiguous items. We aimed for a 30/30/40 percent split in the likely label of the items ('yes', 'no', and 'could be', respectively), allowing for more 'could be' items as we expected to see more structuring within this class.
- *Each task should contain roughly an equivalent amount of structure in the ambiguous class*. This was to reduce the effects of differences in the amount of structure on labeling speed and decision making. Again, this was prompted by our formative study; some participants commented that having too many groups required too many decisions. We focused on structure similarity in the 'could be' class because we expected more structuring and more difficult structuring decisions over ambiguous items.
- *Each task should contain roughly the same number of pairs of items that could reasonably be construed as belonging together*. Examining pairs of items that should belong together (i.e., should have the same label) is our intended mechanism for measuring label quality. An alternative would be to compare the performance of machine learned models built with the labeled data produced by each participant. However, such models are affected by many factors (e.g., concept complexity, feature quality); thus, differences in model performance cannot be entirely attributed to label quality. This is particularly true for small datasets, where a large amount of variance is expected [8]. Therefore, because supervised machine learning is based on the premise that similar inputs should have similar outputs [18], label consistency of pairs of similar items is a reasonable proxy for label quality.

To create these tasks, we again turned to the Open Directory Project. First, we selected candidate concepts meeting our familiarity requirement (e.g., concepts related to everyday activities, such as *cooking* related web pages). Then, two of our experimenters independently coded approximately 160 web pages for each candidate concept. These web pages were selected from a corpus of about 180,000 pages in the Open Directory Project database. Approximately half of the pages coded were listed as part of the concept in the database and half were taken randomly from the rest of the corpus.

The coders applied three high-level codes to the web pages ('yes', 'no', and 'could be') and also created their own groups of 'could be' pages. We then selected web pages that both coders agreed upon, based on the intuition that while people might differ in their labeling decisions on some data, they also might agree on some decisions (e.g., people might disagree about whether a web page about catering services matches the concept of *cooking*, but many people would likely agree that a page containing recipes is about cooking). Selecting items that two people agreed upon therefore helps ensure that the tasks contain data meeting our specified requirements (e.g., having similar proportions of 'yes', 'no' and 'could be', as well as similar amounts of structure).

Throughout this process, we eliminated candidate datasets along the way that broke any of our requirements. Our final set of tasks pertained to the concepts of *cooking*, *travel*, and *gardening*. We obtained 54 manually coded items for each task with 16/16/22 items within the 'yes', 'no', and 'could be' categories, respectively. Each data set also contained six to seven groups within the 'could be' category that both coders agreed upon. We then supplemented this data set with 54 additional items selected by taking the nearest neighbor of each manually coded item (based on cosine similarity over a bag-of-words representation) within our corpus of Open Directory Project pages. Each of our final tasks therefore contained 108 items to label. Note that for our analyses discussed later in this paper, we use all of the data (both the manually coded and supplemental data) to measure labeling speed but use only the manually coded data to measure label consistency in order to ensure that items we considered as pairs should indeed belong together.

Participants and Procedure

Fifteen participants were recruited from a large software company for this experiment (six female) ranging in age from 22 to 45 years old. All participants reported computer use of at least 30-40 hours per week (median of 40-50 hours). No participant had a background in machine learning, and only one worked as a software developer (the rest were program managers or worked in non-development departments, such as marketing and legal).

We used a within-subjects study design, counterbalancing interface order with a Latin-square. Because we did not expect concepts to have carryover effects between tasks, we fixed task order to *cooking*, *travel*, and then *gardening*.

Before each task, participants were given a brief introduction to the interface they would use for that task and time to practice with it. We used the same concept—*libraries*—for the practice tasks. For each actual task, we asked participants to "categorize the web pages in term of whether you think they are about [cooking/gardening/travel] or not", according to their own definition of the target concept.

All interactions with each interface were logged. After completing each task, participants filled out a questionnaire gauging their attitudes toward the interface they used for that task. After completing all three tasks, a final questionnaire was distributed asking participants about their overall preferences for the different interfaces.

RESULTS

Our analyses of the data we collected from our experiment fall into four general categories: tool usage, label quality, labeling speed, and user attitudes and preferences. Unless otherwise noted, we computed quantitative comparisons using the Friedman rank sum test followed by post-hoc Wilcoxon rank sum tests with Bonferroni correction.

Usage of Structured Labeling Supports

Having the ability to structure data does not imply people will actually do so. Because we did not request or require that participants use the structuring supports, we were able to investigate whether their own sense of the usefulness of structuring would outweigh the time and mental effort cost of structuring (according to Attention Investment Theory [6], people will not invest attention in activities unless they think the benefits will outweigh the costs).

If participants did not feel structured labeling was useful, we would expect to see no differences between the numbers of groups (i.e., structurings) across the three conditions. However, we found that participants did indeed make use of structuring supports (Figure 4, left), finishing the study with significantly more groups with the structured labeling conditions than with the baseline (X^2=20.19, df=2, p<.001). Pairwise tests confirmed that both manual (p<.001) and assisted (p<.001) structured labeling resulted in more groups than our baseline (which had three permanent groups). This suggests that participants felt the benefits of structuring their labels outweighed the costs of doing so.

As expected, participants most often structured pages within the ‘could be’ category. Some participants also structured in the ‘yes’ category, but very few structured the ‘no’ category. This was likely because the ‘no’ category contained a wider variety of unrelated pages, making structuring seem less useful or more time consuming. Figure 4 (right) also shows that the ‘could be’ groups were often smaller than ‘yes’ groups, which in turn were often smaller than ‘no’ groups.

Usage of our structured labeling prototypes also revealed evidence of concept evolution. Participants revised their structuring (i.e., moved pages between groups or groups between categories) significantly more often while using the structured labeling conditions than the baseline during the first half of each labeling session (X^2=8.93, df=2, p=.011), with pairwise tests showing more revisions in both manual (p=.006) and assisted (p=.024) compared to the baseline. Interestingly, differences were also observed during the second half of each labeling session (X^2=8.04, df=2, p=.018), however, only the manual condition showed more revisions than baseline during this period (p=.012). These results suggest that structured labeling encouraged concept evolution and the assisted structured supports may have enabled people to solidify their concept definition sooner than with manual structuring alone.

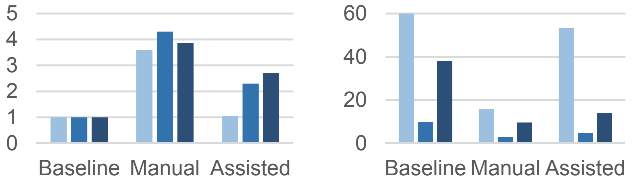


Figure 4. (Left) Average number of groups at the end of the experiment (light=‘no’, medium=‘could be’, dark=‘yes’). (Right) Average number of pages per group (same legend as above).

Figure 5 (right) shows another discrepancy between manual and assisted structuring—participants revisited many more pages using the manual condition than the assisted condition. This was especially pronounced during the second half of the labeling sessions (X^2=12, df=2, p=.002), and pairwise tests confirm that manual significantly differed from baseline (p=.005) and assisted (p=.005) during this time. Again, this discrepancy may be attributable to the assisted structuring supports for recalling group content via summaries (reducing the need to manually review a group’s contents) or for recommending groups (reducing the number of groups created and thereby reducing the number of revisits).

Label Quality

Recall that our mechanism for comparing label quality is to measure label consistency of pairs of items that two independent coders agreed should belong together (described under Conditions and Tasks). We computed label consistency via the Adjusted Rand Index (ARI) [17], a common and recommended metric for computing agreement between some partitioning of data (defined by participant labels in our case) and some ground truth (defined by data our coders agreed upon) by examining pairs [25]. Intuitively, ARI computes the proportion of pairs that *should* (and *should not*) have ended up together (or not) over all possible pairs and is adjusted for chance groupings. We used the ‘could be’ items that our two experimenters independently labeled and agreed upon as our ground truth partitioning (i.e., where pairs grouped together by both annotators were considered similar and those not grouped together were dissimilar). This amounted to 231 pairs to measure ARI over. Note that whether or not a participant considered any item as belonging to their concept is irrelevant to measuring label consistency (and irrelevant to ARI). Instead, we only care that pairs of items that should be together end up together. For example, if a user placed two items together in a group

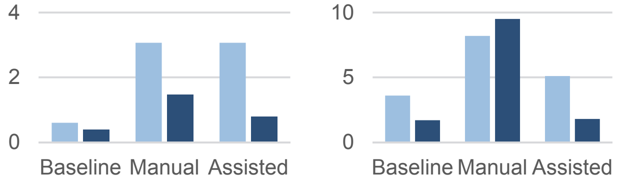


Figure 5. Average number of pages and groups participants revised (left) or revisited (right) during the first half of the experiment (light) and last half (dark).

in the 'yes' category, the ARI measure would still mark them as together, even if our ground truth marked these as together in a group in the 'could be' category).

Our analysis showed a significant effect of interface condition on label consistency according to our ARI metric (X^2=6.53, df=2, *p*<.038). Pairwise tests showed that participants labeled data significantly more consistently in both the manual (*Z*=-2.329, *p*=.02) and assisted structuring (*Z*=-2.329, *p*=.02) conditions than in the baseline condition. No difference was found between the manual and assisted structured labeling conditions (*Z*=-0.852, *p*=.394). These results (shown in Figure 6) suggest that structured labeling did improve the quality of participants' labeled data, helping them to label items in a more consistent manner.

Labeling Speed

To investigate the impact of structuring on labeling speed, we measured the total time it took participants to complete the labeling task in each condition. We found that participants in the baseline condition finished labeling in approximately 10 minutes on average, versus nearly 15 minutes for the structured labeling conditions (X^2=14.93, df=2, *p*<.001). Pairwise tests revealed that the differences between the baseline and manual (*p*=.003) and the baseline and assisted (*p*<.001) structured labeling tools were significant (Figure 7).

We also examined how quickly participants *initially* labeled individual pages with each interface. That is, we measured how long it took participants to examine pages the first time they appeared and decide on an initial label for it. We did not include time they may have spent revisiting that page when refining their concepts. From this analysis, we found a difference in initial label speed between the baseline and manual structured labeling (with baseline being faster), but no difference between the baseline and the assisted structured labeling (X^2=6.40, df=2, *p*=.040; pairwise test between baseline and manual *p*=0.016, no statistically significant differences between the other pairs). These results suggest that our assisted supports might help to mitigate some of the costs of structuring labels.

Additionally, we examined how long it took participants to initially label the first 50% of pages versus the last 50%, with the intuition that people might be able to apply labels faster once their structuring was established. While we found no significant differences in speed from beginning to end, we found that by the end of their tasks the differences in time to label between the baseline and manual and between the baseline and assisted conditions decreased (Figure 8). These changes may be due to a stabilization of participants' structuring as they progressed. Indeed, examining when participants modified their structures, we found that of the 100 pages that were re-examined at least once across all users, 69 were presented during the first half of the task, while only 31 were presented during the second half.

User Attitudes and Preferences

To reveal a larger picture of the impacts of structured labeling, we examined user attitudes and preferences for each interface, as well as their self-reported concept changes.

At the end of the study, we asked participants to rank each tool in order of preference. Figure 9 illustrates an upward trend with more participants ranking the manual structuring tool as their favorite than baseline, and more still ranking the assisted structured labeling tool as their favorite. We also see the opposite trend for participants' least favorite tool, with the baseline tool ranked last most often.

Participants' comments provided some insight as to why they might have preferred the assisted structuring tool. For example, some participants appreciated the group recommendations:

> *"Power of suggestion possibly? It helped with the sorting process but also make you subconsciously say 'wait, that's not right'. It wasn't loud and in your face."*

> *"Assisted grouping was best because you were really 'OK'ing' (or not) what the computer guessed. Otherwise the three simple categories were quickest."*

Similarly, another participant said that without the recommendation feature, he would have preferred the unstructured labeling tool (which he ranked as his second favorite) because it was less complex:

> *"Simple can be useful. Suggestions were extremely helpful. Categories introduced complexity to the system."*

This idea that "simple can be useful" was echoed by other participants, particularly the three who preferred the unstructured labeling tool. However, even among this group, there was an awareness that the assisted structuring tool could be useful in the right circumstances:

> *"I favored the more simple the better [sic] [baseline]… But when I didn't know the topic very well, such as 'Gardening', I was hoping for some assistance in [similar pages] and*

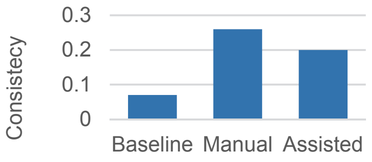


Figure 6. Average consistency (computed by Adjusted Rand Index) of coded pairs. Participants were significantly more consistent while working with structured labeling tools.

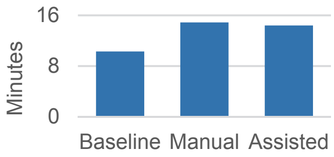


Figure 7. Average duration (in minutes) of the labeling task. Participants finished the task faster using the baseline tool than with structured labeling tools.

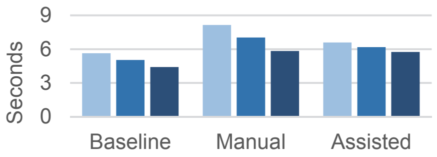


Figure 8. Average length of time (in seconds) it took participants to initially label each item (light=first 54 labels, dark=last 54 labels, medium=average). Participants were slowest with the structured labeling tools.

[group summaries] that can be used to help me categorize."

Another participant also discussed how structuring was particularly helpful when working with unfamiliar topics, including how the tool helped her keep more categories in her mind at once:

"I think I created more groups this time because I'm not as familiar with the topic, so in my mind I wanted to have more categories."

Finally, it is worth noting that participants were aware their concept definitions were changing more often while they were engaged in structured labeling than when using the baseline interface (Figure 10). Friedman rank sum tests showed a significant main effect of interface condition on concept change awareness (X^2=9.91, df=2, p=.007), and pairwise comparisons confirmed a significant difference between baseline and assisted structured labeling (p=.016).

DISCUSSION

We illustrated the concept evolution problem with a series of formative studies, demonstrating that concept evolution impacts people's ability to label data consistently. We then introduced structured labeling as a novel approach to dealing with concept evolution. Our controlled experiment showed that people used and preferred structured labeling over traditional labeling, and that structured labeling improves label consistency but at a cost of speed. However, we also wanted to revisit a finding from one of our formative studies that concept evolution can result in the same people making different labeling decisions on the same data at different times. In particular, we wanted to determine if structured labeling can improve label consistency in this situation.

We conducted a small follow up study with eight machine learning experts. We asked our participants to label 100 web pages from the *gardening* concept and then come back ten days later to re-label the same data (the ordering of the data was shuffled between sessions). Four of our participants used our baseline tool and four used our assisted structured labeling tool. Consistency was computed as in our formative study on participants' high-level categorizations (i.e., 'yes', 'no', 'could be') from the first session to the next.

From this study, we found that structured labeling helped people arrive at more consistent structuring decisions when labeling the same data ten days apart. Participants in the baseline condition showed 86.3% consistency on average (SD=5.3%), while those in the structured labeling condition averaged 90.9% (SD=5.6%). In addition, three of the participants using the baseline interface arrived at significantly different labels after ten days (computed via McNemar-Bowker tests of symmetry), while only one of the structured labeling participant's labels were significantly different from one session to the next. These findings suggest further investigation of the impact of structured labeling on consistency of labeling decisions over time is warranted.

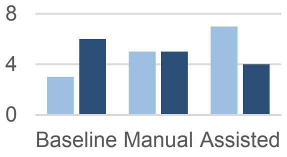


Figure 9. Number of participants who ranked each tool as their favorite (light) and least favorite (dark). Assisted structured labeling had the highest number of favorite and lowest number of least-favorite rankings.

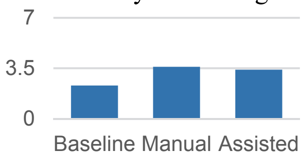


Figure 10. Participant responses to Likert-scale question "How often did your concept change?". Participants reported more changes while using the structured labeling tools.

Binary labeling, however, is not the only potential use for structured labeling. Other tasks that require consistent labels (e.g., multi-class classification, entity extraction) may also benefit, though additional supports for managing more classes or complex inputs may be necessary. Further, the labeling structure itself may be useful to both humans and machines. Labeling guidelines or rules can emerge directly from the structured labeling process, and tools such as our prototype could be used to share these guidelines as a collection of exemplars rather than as written rules. Machine learners may also benefit from this structure; for example, items in certain groups could be reweighted, model selection could explore different combinations of groups, and group-specific features could be identified.

CONCLUSION

This paper introduced the notion of *concept evolution* in machine learning and made the following contributions:

- Results from three formative studies illustrating the impact of concept evolution in machine learning.
- A novel interaction technique for helping people evolve concepts during labeling (*structured labeling*), and two tools instantiating this technique.
- Results from a controlled experiment comparing structured labeling to traditional labeling in machine learning, showing that structuring was used and preferred by participants and helped them label more consistently, but at a cost of speed (particularly early in labeling).
- Results from a follow up experiment comparing label consistency over time, showing that structured labeling helped participants recall their earlier labeling decisions and increased their consistency over time.

Taken together, these results reveal the pervasiveness of the concept evolution problem from machine learning practitioners developing systems for widespread deployment to end users providing training data to their personal classifiers. Structured labeling provides a solution to concept evolution and is a further step on the road toward helping people meaningfully interact with machine learners.

ACKNOWLEDGMENTS

We thank the CHIL team and John Platt at MSR for their feedback and support, and our participants for their time.

REFERENCES

1. Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. Power to the people: The role of humans in interactive

machine learning. *AI Magazine* (under review).

2. Amershi, S., Lee, B., Kapoor, A., Mahajan, R., & Christian, B. CueT: Human-guided fast and accurate network alarm triage. In *Proc. CHI*, ACM (2011), 157–166.
3. Basu, S., Fisher, D., Drucker, S. M., & Lu, H. Assisting users with clustering tasks by combining metric learning and classification. In *Proc. AAAI* (2010), 394–400.
4. Bennett, P. N., Chickering, D. M., & Mityagin, A. Learning consensus opinion: mining data from a labeling game. In *Proc. of WWW* (2009), 121–130.
5. Billsus, D., & Pazzani, M. J. A hybrid user model for news story classification. In *Proc. UM* (1999), 99–108.
6. Blackwell, A. F. First steps in programming: A rationale for attention investment models. In *Proc. HCC*, IEEE (2002), 2–10.
7. Borlund, P. The concept of relevance in IR. *Journal of the American Society for information Science and Technology 54,* 10 (2003), 913–925.
8. Brain, D., & Webb, G. On the effect of data set size on bias and variance in classification learning. In D. Richards, G. Beydoun, A. Hoffmann, & P. Compton (Eds.), *Proc. of the Fourth Australian Knowledge Acquisition Workshop* (1999), 117–128.
9. Brodley, C. E., & Friedl, M. A. Identifying mislabeled training data. *Journal of Artificial Intelligence Research 11* (1999), 131–167.
10. Bshouty, N. H., Eiron, N., & Kushilevitz, E. PAC learning with nasty noise. *Theoretical Computer Science 288,* 2 (2002), 255–275.
11. Carterette, B., Bennett, P. N., Chickering, D. M., & Dumais, S. T. Here or there. *Advances in Information Retrieval* (2008), 16–27.
12. Conway, D., & White, J. M. *Machine Learning for Email: Spam Filtering and Priority Inbox*. O'Reilly (2011).
13. Cunningham, P., Nowlan, N., Delany, S. J., & Haahr, M. A case-based approach to spam filtering that can track concept drift. *The ICCBR 3* (2003).
14. Czerwinski, M., Dumais, S., Robertson, G., Dziadosz, S., Tiernan, S., & Van Dantzich, M. Visualizing implicit queries for information management and retrieval. In *Proc.CHI*, ACM (1999), 560–567.
15. Gabrilovich, E., Dumais, E., & Horvitz, E. NewsJunkie: Providing personalized newsfeeds via analysis of information novelty. In *Proc. WWW* (2004), 482–490.
16. Google. Search quality rating guidelines. *Online:* http://google.com/insidesearch/howsearchworks/assets/searchqualityevaluatorguidelines.pdf (2012).
17. Hubert, L., & Arabie, P. Comparing partitions. *Journal of classification 2,* 1 (1985), 193–218.
18. Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. Supervised machine learning: A review of classification techniques. *Informatica 31* (2007), 249–268.
19. Law, E., Settles, B., & Mitchell, T. Learning to tag using noisy labels. In *Proc. ECML* (2010), 1–29.
20. McGee, M. A look inside Bing's human search rater guidelines. *Online: http://searchengineland.com/bing-search-quality-rating-guidelines-130592* (2012).
21. Paul, S. A., & Morris, M. R. Sensemaking in collaborative web search. *Human–Computer Interaction 26*, 1–2 (2011), 72–122.
22. Rajaraman, A. & Ullman, J. D. "Data Mining". *Mining of Massive Datasets* (2011), 1–17.
23. Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. The cost structure of sensemaking. In *Proc. of INTERACT and CHI*, ACM (1993), 269–276.
24. Robertson, G., Czerwinski, M., Larson, K., Robbins, D. C., Thiel, D., & Van Dantzich, M. Data mountain: using spatial memory for document management. In *Proc. UIST*, ACM (1998), 153–162.
25. Santos, J. M., & Embrechts, M. On the use of the adjusted rand index as a metric for evaluating supervised classification. In *Artificial Neural Networks–ICANN* (2009), 175–184.
26. Sheng, V. S., Provost, F., & Ipeirotis, P. G. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proc. KDD* (2008), 614–622.
27. Stanley, K. O. Learning concept drift with a committee of decision trees. *Tech. Report UT-AI-TR-03-302, University of Texas at Austin* (2003).
28. Teevan, J., Cutrell, E., Fisher, D., Drucker, S. M., Ramos, G., André, P., & Hu, C. Visual snippets: summarizing web pages for search and revisitation. In *Proc. CHI*, ACM (2009), 2023–2032.
29. Tsymbal, A. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin* (2004).
30. Valiant, L. G. Learning disjunctions of conjunctions. In *IJCAI* (1985), 560–566.
31. Westergren, T. The music genome project. *Online: http://pandora.com/mgp* (2007).
32. Whittaker, S., & Hirschberg, J. The character, value, and management of personal paper archives. *ACM TOCHI 8*, 2 (2001), 150–170.
33. Widmer, G., & Kubat, M. Learning in the presence of concept drift and hidden contexts. *Machine learning 23*, 1 (1996), 69–101.
34. Yih, W. & Jiang, N. Similarity models for ad relevance measures. In *MLOAD - NIPS Workshop on online advertisin*g (2010).
35. Yoshii, K., Goto, M., Komatani, K., Ogata, T., & Okuno, H. G. An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model. *IEEE Transactions on Audio, Speech, and Language Processing 16*, 2 (2008), 435–447