Crowdnection: Connecting High-level Concepts with Historical Documents via Crowdsourcing

Nai-Ching Wang
Dept. of Computer Science
Virginia Tech
Blacksburg, VA 24061, USA
naiching@vt.edu

CHI'16 Extended Abstracts, May 07-12, 2016, San Jose, CA, USA
ACM 978-1-4503-4082-3/16/05.
http://dx.doi.org/10.1145/2851581.2890377

Abstract
To form and test hypotheses and finally produce conclusions, people use existing schemas to search a pool of data for evidence. The quality of the search largely depends on the quality of connections between the schemas and the data. Making good connections between schemas and unprocessed data is challenging because it is time-consuming and may require expertise. Crowdsourcing provides a potential solution because with appropriate methods, humans are often more effective at synthesizing diverse information than automated techniques. This paper introduces Crowdnection, which leverages crowdsourcing methods to examine the effect of amount of context on performance in making connections between raw texts of historical textual documents and high-level concepts. The results suggest novices are able to help process information to provide meaningful insights, and indicate that there is an ideal amount of context facilitating the sensemaking process.

Author Keywords
sensemaking; crowdsourcing; text analysis; history.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction

Connecting new data with existing schemas (high-level concepts) is important because one the one hand, people (re)use their existing schemas to search data for evidence supporting or refuting their hypotheses and eventually produce a final conclusion via a top-down approach; on the other hand, people find insights from the new data to better existing schemas via a bottom-up approach [8, 9]. However, there are two big challenges to making strong connections between unprocessed data with existing schemas, e.g. analyzing a textual document for topics or themes it contains. The first challenge is that making connections is time-consuming, especially when there are a lot of unprocessed data. The more time spent on making connections, the less time can be spent on forming and testing hypotheses to produce the final presentation. The second challenge is that connections might be low quality because making appropriate connections often requires certain expertise, e.g. connecting historical primary sources with research topics that historians use. We want connections that can be reused as paths from schemas to data to identify evidence or that can produce suggestions for better schemas.

One possible solution is to use advanced search engines and automated techniques such as topic modeling to help locate and cluster possibly-related information. Using automated techniques may be efficient but recent research shows the quality may not be sufficient [1]. In addition, even with these techniques, human input is usually required to further process the information to make the right connections between new information and existing schemas. This is because topics are often not keywords in the raw texts and a set of keywords might have different connections to different topics based on the context.

Another potential solution is to recruit humans to make connections first and then use computer algorithms to quickly aggregate all the output. On the one hand, we want to lower the initial expertise requirements to recruit as many people as possible to decrease the time for data processing. On the other hand, we want to ensure the quality of connections is good enough. Several crowdsourcing studies have shown that with appropriate methods novices are able to provide output with (close to) expert quality (e.g. [1, 3, 5, 9]). From the literature, we also know that receiving more context is an effective way to help identify structures in the texts, and that should help novices improve their performance. This paper introduces Crowdnection, a system that leverages the above crowdsourcing methods to make crowdsourced connections between historical documents and topics selected by historians so that historians can spend more time focusing on their research questions and/or hypotheses instead of making these connections by themselves. Our research question for this study is "What is the effect of amount of context on the quality of topics chosen by non-expert crowd workers?" Our two quality measures are **agreement with expert** by comparing crowd workers' to experts' choices of topics and **agreement with crowd** by comparing choices of topics among crowd workers.

The results of the study show that novices are able to help process information to provide meaningful insights

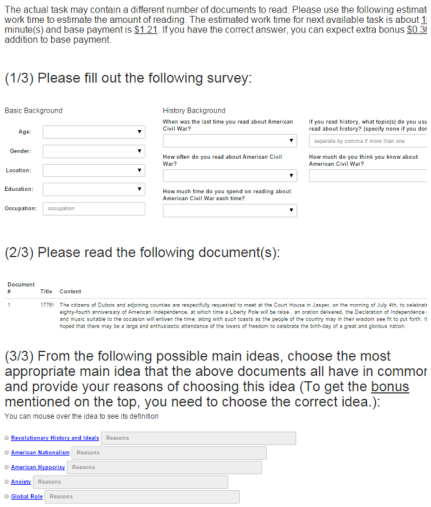

Figure 1: The interface of Crowdnection system.

and suggest that there is an ideal amount of context for this process. This study contributes to the field of HCI by demonstrating: 1) the importance of including a measure to help distinguish errors and surprises in systems that involve sensemaking; 2) the effect of amount of context on the quality of making connections between raw textual data and high-level concepts, which helps identify the right amount of information for various tasks related to learning and data analysis.

Related Work

Techniques from machine learning such as topic modeling provide the ability to group documents and keywords in terms of topics but humans are still required to specify the number of topics in advance and to determine the names of the clusters, that is, the topics. In addition, these algorithms often create incoherent clusters without appropriate supervising [2]. The results from a recent study show that for extracting categories and clusters from complex textual data, around two-thirds of the extracted groupings were not meaningful or interpretable for the dataset used in the study [1].

There are several studies exploring crowdsourcing clustering, summarizing and categorizing textual data with different techniques. For example, Cascade [3] introduced a workflow to generate taxonomies by asking crowd workers to generate categories, select best category and categorize given texts. André et al. [1] proposed a two-stage process (re-representation and iterative clustering) to cluster and elicit categories from texts. And Context Trees [9] presented two-phase (upward and downward) movements through context trees to crowdsource global summarization of a story and a movie with local context. While these studies with textual data focus on crowdsourcing the generation of categories and taxonomies from unprocessed texts, this study focuses on crowdsourcing the generation of connections between existing topics and raw texts.

To synthesize information from diverse sources, Crowdlines [5] finds most effective combination of context and system guidance for merging structured information efficiently. This study differs from Crowdlines in its focus on abstracting topics directly from the texts without meaningful structures, rather than merging information with different existing structures to find the most salient structure. Besides, while most of the prior studies only focus on comparing the crowd's work to experts' work, we also include a measure to help distinguish errors and surprises.

Context Theory of Classification Learning states that more context helps establish a unified language and identify common salient structures from diverse information [6]. But more context also means more time and cognitive resources are required to process the context and that causes fatigue and degradation of performance. Therefore, following the research question, we propose two hypotheses:

H1. Reading more documents increases the agreement of selected topics between crowd workers and experts before reaching some number of documents.

H2. Reading more documents increases the agreement of selected topics among crowd workers in the same condition before reaching some number of documents.

The two hypotheses are related. If H1 is true, H2 is true as well. However, if H2 is true, H1 is not necessarily true because crowd workers may agree with

# of docs	Topic 1	Topic 2	Topic 3	Total
1	33.3%	66.7%	33.3%	44.4%
3	33.3%	100%	33.3%	55.6%
5	0%	100%	0%	33.3%
7	100%	100%	33.3%	77.8%
Total	41.7%	91.7%	25%	52.8%

Table 1: Agreement with Expert of Each Condition Across Different Topics

Topics	***p*-value**
Topic 1	0.1818
Topic 2	1
Topic 3	1
Topic 1+Topic 3	0.1431
ALL	0.3541

Table 2: *p*-value of Fisher's exact test with different topics

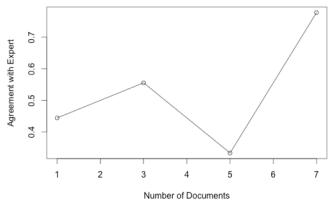

Figure 2: Number of Documents vs. Agreement with Expert

each other but not with an expert. This could be a contradiction (or say, surprise) that usually leads to interesting research [4] and/or a suggestion of the need of new schemas to better fit the documents [8].

Method

Experimental Design

The documents used in the study are historical primary sources such as personal diaries/letters, newspaper reports and public speeches during American Civil War era (roughly 1840-1870). For the common topics, 3 topics were chosen based on similar reading time: "Revolutionary History and Ideas", "American Hypocrisy", and "Anxiety". For the example document shown in Figure 1, the historian-selected topic is "Revolutionary History and Ideas". Definitions of the concepts were provided during the experiment to ensure the common understanding of the concepts. The dataset is a representative example in the sense that 1) it is currently used by historians for their research and 2) the unprocessed texts include rich, sometimes excessive, details without containing any of the topics of interest verbatim. We vary the amount of context by controlling the number of documents the participant sees before choosing the topics. In this study, there are 4 amounts of context corresponding to 1, 3, 5 and 7 documents. Therefore, we have 12 combinations for the 4 conditions and 3 topics in our experiment. For each combination, there is only one common topic across all given documents. We recruited 3 participants for each combination from Amazon Mechanical Turk resulting 36 workers (21 female) in total across all combinations. The crowd workers were required to be from the US and have completed 50 HIT's with 95% approval rate. They were paid minimum wage based on estimated reading time and were told they would get 30% bonus if their answers were the same as the ones chosen by historians as shown on the top of Figure 1.

The quality of the results is measured in terms of agreement with expert and agreement with crowd. The former describes overlap between the participant's concept selection and the historian's, while the latter describes overlap among participants performing the same task. Both measures are important because on the one hand, we want to know if the crowd can produce results comparable to the experts. On the other hand, if there is a difference between experts' and crowd's choices, we would like to know if the difference is merely an error or a valuable surprise. Agreement with expert is the ratio of number of same answers (between crowd workers' and historians') to the number of total crowd workers' answers. Fleiss' kappa and Raw Agreement Indices (RAI) are used for agreement with crowd because we have different crowd workers evaluating different conditions.

Procedure

Once participants accept the HIT on Amazon Mechanical Turk, they first fill out a survey of basic demographics and history knowledge. Then they are directed to read a certain number of documents based on the combination they are assigned to. After reading the given documents, they choose one topic (that is, main idea) from the given list of potential topics which include some other distractor topics along with their reasons of choosing the topic. This task procedure is shown in Figure 1 as three steps.

System Implementation

The Crowdnection system used for the experiment in this study is a customized variant of Incite, a document

# of docs	Fleiss' kappa	RAI
1	0.036	0.333
3	0.333	0.556
5	0.609	0.778
7	0.500	0.667

Table 3: Coefficients of Fleiss' kappa and Raw Agreement Indices (RAI)

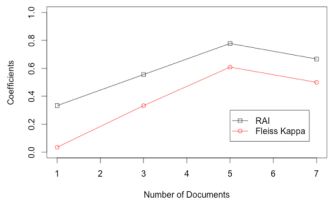

Figure 3: Number of Documents vs. Agreement with Crowd (measured in terms of RAI and Fleiss' Kappa)

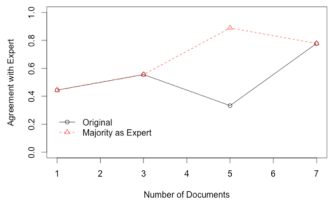

Figure 4: Number of Documents vs. Agreement with Expert (with Majority as Expert)

analysis plugin for a popular online content management system, Omeka. The front-end 1) displays given historical documents in randomized order for a certain combination; 2) collects one topic (common main idea) out of the given list along with reasons, as shown in Figure 1. The back-end stores anonymized user info, user responses and testing data such as documents and selected topics.

Results and Discussion

Agreement with expert

The agreement with expert of all combinations is shown in Table 1. For different numbers of documents, it is lowest at 5 documents and highest at 7 documents. The trend is shown in Figure 2. The agreement with expert for different topics is also different. The agreement with expert of Topic 2 is much higher than the other two concepts. Logistic regression (of # of documents + topics) shows that only Topic 2 has a significant effect on agreement with expert (p=0.02). Since our sample size is small, we use Fisher's exact test to see if there is any significant effect of number of documents on agreement with expert. We run the test with different topics to see overall and individual effects. We add Topic 1+ Topic 3 because we already know Topic 2 has a strong impact on the results. Based on the test results shown in Table 2, there is no significant effect of number of documents on agreement with expert under different combinations of topics. This refutes our first hypothesis.

Agreement with crowd

The agreement with crowd across different numbers of documents is shown in Table 3. For both measures, Fleiss' kappa and raw agreement indices reveal a similar trend illustrated in Figure 3. The agreement with crowd increases as the number of documents increases until the number of documents reaches 5. After that, it decreases as the number of documents continues to increase. Linear regression indicates that there is a significant positive linear relationship between number of documents and agreement with crowd until reaching the maximum, that is, 5 documents with p-values 0.01 (Fleiss' kappa) and 2^{-16} (RAI), respectively. Therefore, the results support our second hypothesis.

Summary and Discussion

The results of agreement with expert show that topics have different saliency of low-level features. Topics with salient low-level features (such as Topic 2 in the study) allow the reader to easily capture the concepts. This might also explain why more documents do not always lead to higher agreement with expert, especially at 5 documents. At 5 documents, from the raw responses, 5 out of 6 disagreed answers are the same topic (Topic 4) that is also the second most common topic among the 5 documents of Topic 1 and Topic 3. We suspect Topic 4 has salient low-level features that attract participants' attention. However, the agreement with expert increase at 7 documents seems to suggest that providing more context can help overcome topics with salient low-level features.

There are some interesting discrepancies between the results of the two measures, especially at 5 documents. The agreement with expert at 5 documents is the lowest while the agreement with crowd at 5 documents is the highest. This suggests crowd workers have a strong agreement against the existing connections of the topics created by historians. This discrepancy might be a "surprise" to historians and the historians might be able to turn the surprise into interesting research [4].

The discrepancy might also suggest the need of new schemas or reorganizing to better fit the documents [8]. This discrepancy also demonstrates the value of including the agreement with crowd measure because with this high agreement with crowd, we have a strong reason to believe this is a surprise instead an error from some individual crowd worker. With the high agreement with crowd, if we see the majority as the expert's answer, the overall agreement with expert at the point will become 88.9%, which is then the highest overall agreement with expert across different numbers of documents as shown in Figure 4.

Conclusion and Future Work

This study shows that in general there is no significant effect of amount of context in terms of number of read documents on the agreement with expert and there is an ideal amount of context for agreement among crowd workers. Moreover, there is a positive linear relationship between number of read documents and agreement before the ideal amount. This study also shows that topics with different saliency of low-level features might affect the ideal amount of context.

Although in general the first hypothesis is refuted, the results seem to suggest we need further investigation with topics and more documents to better understand the effect of amount of context on agreement with expert. In addition, further studies with experts (historians in this study) are also required to better understand the meaning of the discrepancy between the two measures.

Acknowledgements

I would like to thank my advisor, Dr. Kurt Luther for his full support, Michael Stewart for his comments on the draft, Jing Luo for her help of the poster and the Mapping the Fourth team for the dataset and Incite platform. This research is supported in part by the National Historical Publications & Records Commission.

References

1. Paul André, Aniket Kittur, and Steven P. Dow. 2014. Crowd Synthesis: Extracting Categories and Clusters from Complex Data. Proceedings of the 17th ACM CSCW, 989–998.
2. David M. Blei. 2012. Probabilistic Topic Models. Commun. ACM 55, 4: 77–84.
3. Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. 2013. Cascade: Crowdsourcing Taxonomy Creation. Proceedings of the SIGCHI CHI, 1999–2008.
4. Davis, M. S. (1971). That's interesting. Philosophy of the social sciences, 1(2), 309.
5. Kurt Luther, Nathan Hahn, Steven P. Dow, and Aniket Kittur. "Crowdlines: Supporting Synthesis of Diverse Information Sources through Crowdsourced Outlines." In *Third AAAI HCOMP*. 2015.
6. Medin, D. L., and Schaffer, M. M. Context theory of classification learning. Psychological review 85, 3 (1978), 207.
7. Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. Proceedings of international conference on intelligence analysis, 2–4.
8. Daniel M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. "The Cost Structure of Sensemaking." In *Proceedings of CHI '93*, 269–76.
9. Vasilis Verroios, and Michael S. Bernstein. "Context Trees: Crowdsourcing Global Understanding from Local Views." In *Second AAAI HCOMP*, 2014.