Crowd-sourcing Web Knowledge for Metadata Extraction

Zhaohui Wu[†], Wenyi Huang[‡], Chen Liang[‡], C. Lee Giles[‡†]
[†]Computer Science and Engineering, [‡]Information Sciences and Technology
Pennsylvania State University, University Park, PA 16802, USA
zzw109@psu.edu, {wzh112,cul226,giles}@ist.psu.edu

ABSTRACT

We explore a new metadata extraction framework without human annotators with the ground truth harvested from Web. A new training sample is selected based on not only the uncertainty and representativeness in the unlabeled pool, but also on its availability and credibility in Web knowledge bases. We construct a dataset of 4329 books with valid metadata and evaluate our approach using 5 Web book databases as oracles. Empirical results demonstrate its effectiveness and efficiency.

1. INTRODUCTION

In most machine learning settings, unlabeled data is usually cheap and copious, while manually harvesting high quality labels can be costly. Instead of simply labeling all the data or a randomly sampled subset, active learning attempts to select unlabeled data for labeling in a smarter way in order to gain better learning accuracy with lower cost [5]. A key assumption of active learning is that there is a single omniscient oracle that can always correctly tell the ground truth with zero or constant cost. However, this does not hold in real world situations since expert labelers are scarce and expensive and human annotators unavoidably produce errors, often to the distraction of quality in their labeling. To address these issues, recent research has shifted to investigate more practical active learning scenarios based on multiple non-expert labelers and Internet-based crowd-sourcing [3, 4, 10]. With this paradigm, there is no omniscient oracle; instead, multiple imperfect labelers with varying expertise usually produce uneven, inconsistent, subjective and unreliable labels.

Dependence on human labelers can still limit the feasibility of active learning in real world applications. Crowd-sourcing labelers are cheap non-experts but they often produce noisy labels. Moreover, human labelers in crowd-sourcing environments are not always online or available "on demand". Given a search engine that indexes metadata for scientific papers and books in various domains, its crawler can crawl hundreds of thousands new documents to be ingested. The active learner may select less than 100 new documents to be labeled. Even though, with crowd-sourcing tools such as Amazon's Mechanical Turk, there is no efficient way to guarantee that the labeling task will be finished within a fixed time due to the uncertainty of human performance.

We observe that the Web per se provides abundant knowledge bases that can be helpful for obtaining labels for machine learning tasks. For example, given an ISBN of a book, Google book, Amazon book, AbeBooks, etc. can provide metadata information for the book. We thus can consider each knowledge base as an oracle that can provide labeling information. Compared with the active learning in traditional human labeler settings, the web knowledge based approach differs in the following ways.

Communication with oracles: 1) the active learner needs to make well designed queries to get meaningful feedback from the oracles; 2) the active learner might need to do further extraction and inference based on the feedback; 3) batch-mode is enabled where multiple queries can be conducted in a parallel, distributed style since the communication/interaction between active learner and oracles is much more convenient and efficient.

Query strategy: 1) the oracles can provide helpful information for selecting new training instances (for example, an instance with more reliable feedback should be favored), so new training instances selection strategy should be based on not only the informativeness or representativeness in the local unlabeled pool, perhaps, but also the availability and credibility in the oracles; 2) the oracles, always staying online and available "on demand", perform consistently over time, making it easier to model their quality or credibility in the first place.

Cost model and stop criteria: 1) labeling costs are nearly constant across different instances, mainly consisting of query submission and processing, ground truth extraction and inference, which could be accurately measured by time; 2) stop criteria might be application dependent, but the capacity limit of an oracle may force a stopping (for example, Google book allows only a limited number of queries per day per user/IP).

All the above could create a new system with more potential for real world applications due to easy-to-control oracles and cheaper annotation cost. The key consequential questions include: how to select new training instances based on both local and web knowledge? And how to effectively harvest the ground truth from the feedback of the oracles? The first one is the essential query strategy problem in active learning while the second could be regarded similarly as a truth discovery problem [11, 7].

This work has three main contributions: first, we present a novel metadata extraction framework using Web knowledge bases as the oracles. Besides, we propose a hybrid pool-based query strategy based on not only the informativeness and representativeness of unlabeled instances in the local pool, but also the availability and credibility in the Web knowledge bases. Finally, we apply this new framework to a real world application book metadata extraction.

978-1-4799-5569-5/14/$31.00

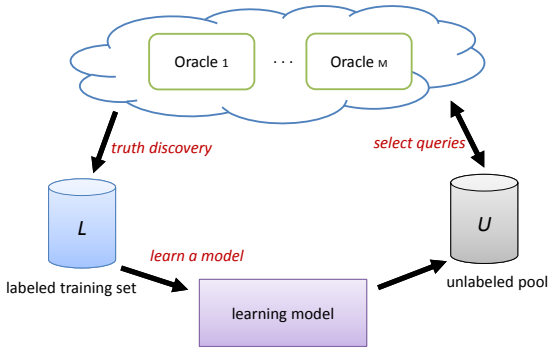


Figure 1: Web knowledge based active learning cycle

2. METHODS

The overall cycle is shown in Figure 1. The active learner incrementally selects new instances to train a more accurate model with a minimal labeling cost until reaching the stop criterion. There are three main disparities from traditional active learning. First, the oracles are not human annotators but web knowledge bases, making the query more efficient. Second, query selection is based on both local and web knowledge (indicated by a double arrow). Third, the truth discovery process is added to find the confident true labels.

2.1 Query Strategy

A commonly used query strategy is uncertainty sampling, which queries the instance whose prediction is the least confident on the trained model, or with a class posterior closest to 0.5 [6]. Since we consider the batch-mode which allows the learner to query instance in groups, a naive strategy would be greedily selecting the top K instances based on uncertainty sampling. However, this strategy fails to consider the "overlap" among those top instances. To incorporate diversity and density measures to encode the "representativeness" of the selected instances on the whole data, we propose a new batch-mode active learning strategy derived from information density [8].

$$\Phi(x) = \phi_i(x)\big(\lambda\phi_{de}(x) + (1-\lambda)\phi_{di}(x)\big)^{\beta}\phi_a(x)\phi_r(x) \quad (1)$$

where $\phi_i(x)$ measures the "base" informativeness to which the uncertainty-based or other metrics in the literature of active learning can be applied; $\phi_{de}(x)$ measures an instance's average similarity (density) to the unlabeled pool U, as shown in Eq. (2); $\phi_{di}(x)$ measures an instance's average dissimilarity (diversity) to the current batch Q, as shown in Eq. (3); λ is a parameter controlling the tradeoff between density and diversity and the weighted sum reflects the representativeness of an instance; β controls the relative importance of representativeness; $\phi_a(x)$ is a boolean function indicating the availability of an instance on all the oracles; $\phi_r(x)$ measures the reliability of the ground truth.

$$\phi_{de}(x) = \frac{1}{|U|}\sum_{x_u \in U} \text{sim}(x, x_u) \quad (2)$$

$$\phi_{di}(x) = \frac{1}{|Q|}\sum_{x_q \in Q} \text{diff}(x, x_q) \quad (3)$$

2.1.1 *Informativeness*

The informativeness is encoded into our query selection strategy by $\phi_i(x)$ based on uncertainty sampling. For a multi-class classification or sequence labeling problem, we use the least confident:

$$\phi_i(x) = 1 - P(y^*|x;\theta) \quad (4)$$

where y^* is the class label with the highest posterior probability under the model θ, or the most likely label sequence (the Viterbi parse). This metric can also be interpreted as the expected 0/1-loss, i.e., the model's belief that it will mislabel an instance [2]. In our experiments on book metadata extraction, we can directly compute this metric if treating it as sequence labeling problem using sequence models such as CRF. If we treat it as multi-class classification problem using SVM, then we compute $P(y^*|x;\theta)$ as the probability product of all predictions in an instance, i.e. $\Pi_i P(y_i^*|x_i;\theta)$, where y_i^* is the most likely label of the token x_i (a line/block in the title page of a book) in the instance x (a book).

2.1.2 *Representativeness*

The uncertainty-based strategies are prone to outliers, or the least certain instances on the classification boundary, but fail to consider those that are representative of the underlying distribution. To this end, the representativeness is encoded via two facets: the density in the unlabeled pool and the diversity in the current batch. The sim and diff computes the similarity and difference between instances respectively. If the similarity function ranges between zero to one, we can simply define diff $= 1 -$ sim. We choose cosine similarity since imperial comparisons have shown it is the clear winner in sequence labeling and text classification corpora [8].

2.1.3 *Availability and Reliability*

It is hard to tell if a human labeler is capable of labeling an instance or not unless such an option is provided and the labeler is unbiased and conscientious. However, in the knowledge bases environment, this can be easily found out based on the feedback. For example, the knowledge bases will return a empty or "not found" message. The availability is an indicator to show whether an instance the active learner want to query is accessible in the oracles.

$$\phi_a(x) = \begin{cases} 0 & \text{if no legal feedback or } \phi_r(x) < \delta \\ 1 & \text{otherwise} \end{cases} \quad (5)$$

Suppose $Y(x)$ denotes the set of returned answers and $f(y)$ denotes the confidence of each answer $y \in Y$, then a simple strategy is to choose the answer with the largest confidence, or $\phi_r(x) = \max\{f(y) : y \in Y(x)\}$. However, this could be problematic if there are multiple maximums. Hence we define it as $\phi_r(x) = f(y_1^*) - f(y_2^*)$ where y_1^* and y_2^* are the first and second most confident answers in $Y(x)$, respectively. If multiple maximums exists, $\phi_r(x)$ will be zero thus no answer will be selected as the ground truth. Otherwise, if $f(y_1^*) - f(y_2^*) \geq \delta$, y_1^* will be selected as the final ground truth for x with reliability $\phi_r(x)$. A simple case is that there is only one candidate answer y in Y. Then y will be directly selected as the truth with reliability $f(y)$.

2.2 Ground Truth Discovery

Obtaining the ground truth from all the feedback returned by the oracles can be regarded as a truth discovery or fact finding problem, i.e. to find out the true facts from a large amount of conflicting claims provided by multiple sources. But the key difference here is that knowledge bases provide much more independent and reliable information than does general sources, with less conflicting claims among them. For example, by manually checking 100 books' metadata returned by Google book, Amazon book, AbeBooks, ISBN Search, and BookFinder4U, we find more than 90% of them agree with each other on title and authors and 100% agree on ISBNs. And the disagreement is usually due to abbreviation difference (e.g. Michael Jordan, M. Jordan, Jordan M.), or omit of the last authors. Based on this observation, it is reasonable to assume that the answers from two different oracles are independent with each

Table 1: Features for Book Metadata Extraction

Feature	Description	Type
font size	*Initial Font*: the font size of the starting character	float
	Average Font: the average font size of all the characters	float
	Font Changes: number of changes in font size	int
location	*Start X*, *End X*, *Start Y*, *End Y*: the coordinates of the line block in the page	float
	Line Number: the (order) number of the line within the page, e.g. 2 indicates the second line	int
	Page Number: the (order) number of the page	int
text	*Bag-of-word*: Top 200 words selected by DF rank in the whole dataset; 1 indicates a word is in the line	boolean
others	*#Words*: total number of words in the line	int
	#Digits: total number of digital words in the line	int

other. Suppose there are M oracles $O = \{o_i\}$ and each has a credibility $c(o_i)$, or the probability that o_i will tell the truth. Moreover, we assume $c(o_i)$ has a uniform distribution on all answers o_i provides. For an answer y returned by a subset $S \subseteq O$, the confidence of y can be calculated by

$$f(y) = 1 - \Pi_{o \in S}\big(1 - c(o)\big) \tag{6}$$

The credibility $c(o)$ can be initialized based on prior knowledge and then updated based on the oracle's performance each learning cycle.

$$c_t(o) = c_{t-1}(o) + \varepsilon \frac{\sum_{x \in Q} \mu(y_x) f(y_x)}{|Q|} \tag{7}$$

where y_x is the answer returned by oracle o for query x and $\mu(y_x)$ indicates whether y_x is selected as the true answer, i.e.

$$\mu(y_x) = \begin{cases} 1 & \text{if } y_x = \text{argmax}_{y \in Y(x)} f(y) \\ -1 & \text{otherwise} \end{cases} \tag{8}$$

2.3 Considering Real Cost

If assuming the labeling cost of an instance is a constant, then the cost can be simply measured using number of labeled instances. But this assumption can hardly hold in human annotation environment. First, labeling tasks could be arbitrarily complex, making the required time varies substantially. Second, the quality of annotators in terms of expertise and responsibility, might also varies considerably, making the actual annotation costs differ from one to another. This motivates the cost-sensitive active learning which tries to minimize the overall cost rather than the number of labeled instances [9, 3]. However, in the web knowledge based active learning, although different oracles may have different response time for a query, labeling costs are nearly constant across different queries for a single oracle. The cost of a query is define as:

$$cost(x) = \sum_{o \in S} t_q(x, o) + t_{td}(x) \tag{9}$$

where S is a subset of all oracles, $t_q(x, o)$ is the time for querying x on oracle o, and $t_{td}(x)$ is the time for harvesting ground truth. Luckily, both $t_q(x, o)$ and $t_{td}(x)$ can be well estimated using the average by issuing a few queries beforehand. If the query strategy is to adaptively select different oracles for a learning cycle or a particular instance, then the cost will vary due to the difference of S. We can now use the benefit-cost ratio (BCR), or

$$\Phi_{BCR}(x) = \Phi x / cost(x) \tag{10}$$

In practice, if we do not use a large number of oracles, we can simply query all oracles every time or set $S = O$ and then the cost will be a constant.

3. EXPERIMENTS

The initial book dataset contains 237,429 pdf documents collected from Citeseerx repository, identified using simple rules "has an ISBN in the first 4 pages" and "number of pages>=100". We then put all the extracted ISBNs to query metadata from Google book (which dominates all book oracles in quality and scale). Those queries together give us 4329 books with valid metadata (title and authors). The first four pages of these books are then extracted line by line and represented by computational features shown by Table 1. The lines with text content matched to the title and authors are labeled by '1' and '2' respectively; others are labeled by '0'. By ruling out those not being successfully extracted or perfectly matched, we finally have 2496 books with ground truth.

We use Libsvm[1] to train a 3-class model using the default parameters. The oracles we use include Abebooks, Amazon book, ISBNSearch, and BookFinder4U (Table 2). We rule out Google book as it has been used to generate the ground truth. The confidence score of each oracle is defined as the groundtruthed rate of its response, or the ratio of number of books with ground truth to the number of books with valid feedback. All experiments were conducted on a machine with 2.35GHz Intel(R) Xeon(R) 4 processors, 23GB of RAM,and Red Hat Enterprise Linux Server(5.7) installed.

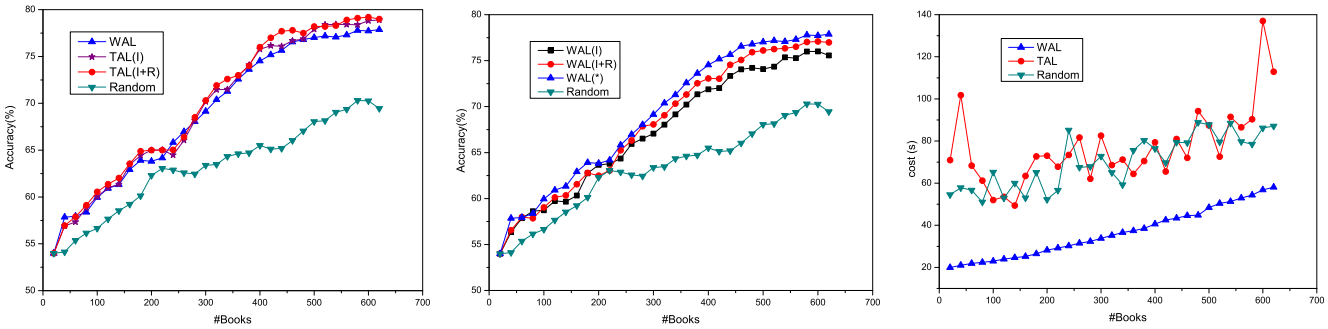
3.1 Evaluation Results

The first question we need to answer is "can the selected oracles give valid labeling?" We queried all the ISBNs of the 2496 books from the four websites, and found that there is only one book that cannot get valid ground truth (ISBN = 9292230522). The number of recalled books and the groundtruthed books (books with true ground truth) from each oracle are shown in Table 2, and the unification of all the groundtruthed books covers 2495 books. This guarantees that at least one oracle will give the correct answer for any query from the 2495 books. In the beginning of the learning, we set the credibility of all the oracles ($c(o)$)equally to be 0.95. Hence, according to Equation (6), books with valid feedback from more oracles will have a larger confidence. The final confidence of each oracle is shown in the last column of Table 2.

Now we check if the active learner is comparable to that trained by a true oracle (based on our ground truth). We show the learning curves of the web oracles based active learner (WAL) compared with those of the true oracle (TAL) in Figure 2a. WAL considers all informativeness, representativeness, availability and reliability for query selection. The batch (Q) size is set to be 20. $\lambda = 0.5$ and $\beta = 1$. TAL(I) denotes the traditional active learning using informativeness, where the most 20 uncertain instances are added to previous training set to train a new active learner. TAL(I+R) denotes the one using both informativeness and representativeness. It is important to note that here a training instance is not a book but an extracted "line" or visual block within it. A page could have more than 30 lines, but with few '1' and '2' lines. To make it a balance classification task, we remove the lines whose previous and next line are both '0'. This will make a book with near equal number of '0', '1', and '2' lines and the total number of lines range from 2 to 20 (the total number of lines for all the 2496 books are 3473). So, the informativeness and representativeness of a book are actually the average among all the lines. Random denotes the one using random selection strategy. For each method, the first 20 books are randomly selected; the learning curve is the average results of 10 runnings. We use a separate test set containing 1000 books. The accuracy is the average results of all the tested lines. WAL shows a very close learning curve to TAL(I) and TAL(I+R) and significantly outperforms the Random method. The book-level accuracy (a book is correctly labeled if all its lines are correctly

Table 2: Oracles for Book Metadata Extraction

Name	#Recalled books	#Groundtruthed books	Query method	Query time (s)	Confidence
Google book	-	-	https://www.googleapis.com/books/v1/volumes?q=isbn:ISBN	0.23	100%
Abebooks	814	777	www.abebooks.com/servlet/SearchResults?isbn=ISBN	0.77	95.5%
Amazon book	1275	1219	www.amazon.com/gp/search/ref=sr_adv_b/?field-isbn=ISBN	0.82	95.6%
ISBNSearch	1170	1118	http://www.isbnsearch.org/isbn/ISBN	1.9	95.6%
BookFinder4U	1301	1237	www.bookfinder4u.com/IsbnSearch.aspx?mode=direct&isbn=ISBN	2.8	95.1%



(a) Learning curves comparisons (b) WAL in different query strategies (c) Cost curves comparisons

Figure 2: Learning and cost curves comparisons

labeled) decreases around 30% for all the methods but the trends of the learning curve remain.

We also compared the performance of different query selection strategies within our framework, showing in Figure 2b. WAL(*) uses the whole query strategy while WAL(I) uses only informativeness and WAL(I+R) uses informativeness together with representativeness. We can see that WAL(*) outperforms all the other three methods due to its more solid active query selection.

3.2 Run-time and Scalability

The most important advantage of the WAL over TAL is its efficiency and easy-to-control. The average query time for a book is only 5 seconds (parallel queries will reduce to 3s), which cannot be achieved by human labelers. The truth discovery time $t_t d(x)$ include the time for extracting the metadata from the resulting content (typically using regular expression) and computing $f(y)$ in Eq.(6), costing only 0.6s in average. The feature extraction and computation costs 13s. The training and testing time using Libsvm ranges from 1 to 38 seconds, positively related to the size of training set (ranging from 20 to 620).

We plot the corresponding cost curve of the learning curves of WAL and TAL, shown in Figure 2c. The cost in terms of time represents all the whole time cost in each learning cycle. Other variations of them are not shown since they are every close to them respectively. Since we do not really have human labelers for testing, in both TAL and Random, we assume the labeling time is proportional to the number of lines of all queried books (5 lines per second), without any delay, which in practice it is hard to achieve by multiple human labelers. The results clearly indicate that our approach is much more efficient and stable in time cost.

4. CONCLUSION AND FUTURE WORK

The inherent uncertainty and noise in human annotation makes crowd based active learning not a practical choice for large scale learning tasks. The recent emerging of high quality and large web knowledge bases provide an alternative approach using oracles for some learning tasks such as information extraction or entity recognition. We proposed this new direction for efficient large scale machine learning whose ground truth can be reliably obtained from web knowledge bases, without any human labeling involved. Our experiments demonstrate its potential for some real world applications where their required "labels" can be harvested from Web. In the future, we will apply the technique to larger datasets of academic books and papers and other applications.

5. ACKNOWLEDGEMENTS

We gratefully acknowledge partial support from the National Science Foundation.

6. REFERENCES

[1] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[2] A. Culotta and A. McCallum, "Reducing labeling effort for stuctured prediction tasks," in *Proceedings of AAAI*, 2005, pp. 746–751.

[3] P. Donmez and J. Carbonell, "Proactive learning: Cost-sensitive active learning with multiple imperfect oracles," in *Proceedings of CIKM*, 2008, pp. 619–628.

[4] M. Fang, X. Zhu, B. Li, W. Ding, and X. Wu, "Self-taught active learning from crowds," in *Proceedings of ICDM*, 2012.

[5] Y. Guo and R. Greiner, "Optimistic active learning using mutual information," in *Proceedings of IJCAI*, 2007, pp. 823–829.

[6] D. Lewis and W. Gale, "A sequential algorithm for training text classifiers," in *Proceedings of SIGIR*, 1994, pp. 3–12.

[7] J. Pasternack and D. Roth, "Making better informed trust decisions with generalized fact-finding," in *Proceedings of IJCAI*, 2011, pp. 2324–2329.

[8] B. Settles, "Curious machines: Active learning with structured instances," in *PhD thesis*, 2008.

[9] B. Settles, M. Craven, and L. Friedland, "Active learning with real annotation costs," in *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, 2008.

[10] Y. Yan, R. Rosales, and G. Fung, "Active learning from crowds," in *Proceedings of ICML*, 2011, pp. 1161–1168.

[11] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," in *Proceedings of KDD*, 2007, pp. 1048–1052.