Active Surveying

Hossam Sharara
Computer Science Department
University of Maryland
College Park, MD
hossam@cs.umd.edu

Lise Getoor
Computer Science Department
University of Maryland
College Park, MD
getoor@cs.umd.edu

Myra Norton
Community Analytics
Baltimore, MD
mnorton@comlytics.com

Abstract

Opinion leaders play an important role in influencing people's opinions, actions and behaviors. Although a number of methods have been proposed in literature for identifying influentials using secondary sources of information, the true set of opinion leaders revealed through primary sources, such as surveys, is still favored in many domains. In this work, we present a new active learning method, which combines secondary data with partial knowledge from primary sources, to guide the primary information gathering process. We apply our proposed active surveying method to a dataset from the pharmaceutical domain, and show how we are able to efficiently obtain the true set of opinion leaders while minimizing the required acquisition of primary data.

1 Introduction

Studying influence in social networks is an important topic that has attracted the attention of a variety of researchers [1–4], as people usually seek the opinion and advice of their peers regarding various decisions, whether it is to try a new restaurant, buy a certain product or even to support a particular politician [5]. This behavior gives rise to certain set of individuals in the social network, referred to as *influentials* or *opinion leaders*, who have a huge impact on other people's opinions, actions and behavior.

In the commercial space, the question of how to identify true opinion leaders within a given population of purchasers or decision makers is of great importance [6–8]. Identifying these individuals properly leads to more effective and efficient sales and marketing initiatives [9]. This is true in multiple industries, but we begin our exploration in the pharmaceutical space, studying the influence networks of local physicians relative to the treatment of specific disease states, which has been the focus of multiple studies in the health care literature [10, 11].

Methods for identifying opinion leaders can be classified into two categories; methods based on primary data or methods based on secondary data. Primary methods rely on manually collecting information about peer-influence in a given population from the individuals themselves. One of the most commonly used primary methods is surveys, where the respondents are asked to report their opinion about who they perceive as opinion leaders. Although primary methods are considered to be the most informative about actual peer-influence, their main drawback is the high costs due to the manual nature of the process: in many cases the surveys are obtained through one-on-one interviews with the respondents, sometimes over the phone, but in many cases in person.

On the other hand, secondary methods rely mainly on using an underlying interaction network as a "*proxy*" for influence, thus avoiding the manual aspect of the primary methods. The most commonly used technique in this setting uses network centrality measures on these secondary networks (e.g., co- authorship, citation, etc...) to distinguish the opinion leaders. However, the correlation between peer-influence in the actual social network and the interactions occurring in the proxy networks cannot be verified. In a recent study about public opinion formation [12], the authors showed through a series of experiments that the customers who are critical in accelerating the speed of diffusion need not necessarily be the best connected.

In this work, we show how to combine the use of primary and secondary methods for leadership identification in the pharmaceutical industry. We study primary data describing a physician nomination network, in which physicians provide survey information describing whose opinion they trust and who they turn to for advice about treating different disease stages. We view this data together with secondary data describing publication history (co-authorship and co-citation), as well as hospital affiliation information. We use ideas from active learning literature to build a model that is able to use partial knowledge of nomination data, together with secondary data, to cost effectively target additional primary data collection via surveys.

The results of this work provide a model for minimizing the amount of primary data needed for accurate leadership identification. As this type of primary data collection typically requires significant investment, this finding empowers organizations to tackle the task of accurate leadership identification in a much more cost effective and efficient manner.

2 Background

Often, secondary data describing assumed influence is easy to obtain, whereas primary data that represents direct relationships of trust and advice-seeking is harder and much more expensive to get. For instance, obtaining a co-authorship network between a set of authors in a certain field (e.g. infectious disease) can be constructed easily by looking at the publication record of each author. However, identifying the influence of each author in the same set requires additional information, and often involves a labor-intensive process of interviewing subjects and extracting their 'network of influence'; who they turn to for advice and recommendations.

In this work, we build on ideas from the field of active learning [13] where the learner is able to acquire labels of additional examples and the goal is to construct an accurate classifier while minimizing number of labeled examples acquired. This is achieved by providing an intelligent, adaptive querying technique for obtaining new labels to attain a certain level of accuracy with minimal training instances. A generic algorithm for active learning is described in [14], where a learner is applied to an initial sample L of labeled examples, then each example in the remaining unlabeled pool UL is assigned an "*effectiveness score*", based on which the subsequent set of examples to be labeled is chosen from UL until some predefined condition is met. The main difference between various active learning methods is how to compute the effectiveness score of each example, which usually corresponds to the utility that it can add to the learning process.

One widely used method for active learning is uncertainty sampling [15], where the learner chooses the most uncertain data point to query, given the model and parameters. Measuring the uncertainty depends on the underlying classifier used, but it is usually translated to how close the data point is to the decision boundary. For instance, if a probabilistic classifier is used, the posterior probability $P(C_i|x)$ can be used directly to guide the selection process. By acquiring the labels for the data points closer to the decision boundary, the model can be improved as a result of narrowing the existing margin. A variety of active learning methods have been proposed [13], with various ways to reduce the generalization error of the underlying model during learning. Active learning has proved to be particularly useful in settings where acquiring labeled data is expensive. It has been applied successfully in various domains, such as speech recognition, image processing, health care, and text analysis.

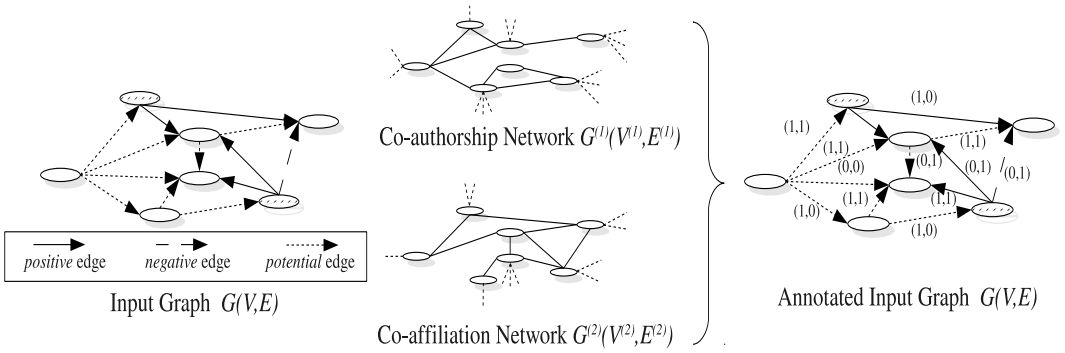

Figure 1: Feature generation for an example author network

3 Active Survey

In this work, we propose an active survey model which applies ideas from active learning to the problem of intelligently gathering primary data for opinion leader identification. We formulate our problem as finding the minimal set of respondents needed to correctly identify at least k% of opinion leaders present in the network. In order to achieve our goal, we need a method that is able to combine partial knowledge from primary sources, such as available user surveys, with other data available from secondary sources to select the next survey respondent such that the set of currently identified opinion leaders is maximized. We use a simple threshold model for identifying opinion leaders; if a candidate receives more than α nominations, he/she is considered an opinion leader.

A key difference in this problem setting is that the acquisition of a survey response is more complex than the acquisition of a single label in traditional active learning. A survey response is a structured object that includes a set of nominations. In some cases there may be weights associated with the nominations; here we are assuming equal weights, although it would be straightforward to extend the model to cases where weights do vary.

Our proposed method relies on providing a dynamic, active learning setting for guiding the survey process. By utilizing a probabilistic classifier, the next survey respondent is chosen to maximize the likelihood of identifying new opinion leaders as a result of her nominations. After the proposed respondent is surveyed, the survey results are fed back to the model to adjust future predictions.

First, we need to define the conditions upon which the next respondent should be selected in the optimal settings. Suppose we are given an initial set of survey responses, and a threshold α that determines the number of nominations an individual should obtain to be declared as an opinion leader. From the initial set of responses, we can generate the following two sets of individuals:

$$\begin{aligned} leaders \quad & \{l : |nominate(v,l)| \geq \alpha; \quad v \in respondents\} \\ candidates \quad & \{c : 0 < |nominate(v,c)| < \alpha; \quad v \in respondents\} \end{aligned}$$

where the *leaders* set represents the set of individuals who have received at least α nominations and are already identified as opinion leaders, where as *candidates* are the set of individuals who have been nominated by at least one person, but have not yet received the required number of nominations to be declared as opinion leaders.

Ideally, we prefer surveying a respondent who is more likely to nominate new leaders; either from the ones already in the candidates set or introduce new individuals to expand it, and less likely to nominate individuals in the already identified leaders set. By following a greedy approach based on this criterion in surveying possible respondents, we can maximize the set of identified opinion leaders with minimal number of respondents.

However, because we don't know in advance the resulting nominations of a given survey, we need to model the responses based on existing secondary source, such as different types of user-interaction networks, along with primary information from current surveys. Then, we can use that model to predict future nominations within an acceptable level of accuracy.

3.1 Model Representation

The input data can be represented as a graph $G(V, E)$, where a node $v \in V$ in the network represents an individual in the population, and the directed edge $e(u, v) \in E$ indicates that v is a possible candidate for respondent u. Thus, the input graph can be viewed as the set of potential nominations in the network.

Generally, the set of potential edges in the network can be as large as $|V| \times |V|$, yielding a fully connected graph. However, in real scenarios, the number of potential edges can often be limited by using the appropriate filters on the incident nodes. We refer to the set of edges in the graph that correspond to the actual respondent nominations as "positive" edges, and the ones that are not realized through the survey as "negative" edges. We refer to the set of edges corresponding to the initial set of surveys as the "observed" edges.

The secondary sources of information are represented in our model as: a) a set of features $\mathcal{F}_v$ associated with the nodes V in G, and b) a set of secondary networks $G^{(1)}(V^{(1)}, E^{(1)}) \dots G^{(n)}(V^{(n)}, E^{(n)})$ representing other types of interactions between individuals (e.g. communication, co-authorship, co-affiliation, etc.). As these secondary networks may not necessarily align with the main graph G; we only consider the sub-graphs $\{G'^{(x)}(V'^{(x)}, E'^{(x)}) : (V'^{(x)} \subset V), (E'^{(x)} \subset E)\}$ that overlap with our network of concern. After aligning the resulting sub-graphs with the input graph G, another set of edge features $\mathcal{F}_e$ are generated for the set of edges E in G, each representing an indicator of the existence of the associated edge in the corresponding secondary network. During this step, the set of node features $\mathcal{F}_v$ are also enriched by additional features from the secondary networks.

3.2 Likelihood Estimation

We can view the problem of identifying opinion leaders as a probabilistic inference problem for unobserved nominations, where the evidence includes the observed nominations along with their associated feature values. Given the input graph G and the sets of node features $\mathcal{F}_v$ and edge features $\mathcal{F}_e$, a probabilistic classifier C is trained using the initial set of observed edges. The trained classifier is then used to infer the label of the remaining unobserved edges in G. For each unobserved, potential edge $e(u, v) \in E$, the classifier outputs a posterior probability of that edge being positive, denoted as $p(+|e(u, v))$, or negative, denoted as $p(-|e(u, v))$.

Assuming that the respondents' choices of influential peers are independent, and given the output probabilities from the classifier along with the initial sets of *leaders* and *candidates* determined by the observed edges in G, we define a likelihood function $L(v)$ for each node $v \in V$ as:

$$L(v) = \frac{\prod_{u_y \in leaders} p(-|e(v, u_y))}{\prod_{u_x \in candidates} p(-|e(v, u_x))}.$$

To avoid numerical instabilities and multiplication underflow, we use the log-likelihood instead

$$\mathcal{L}(v) = \sum_{u_y \in leaders} log(p(-|e(v, u_y))) - \sum_{u_x \in candidates} log(p(-|e(v, u_x))) \quad (1)$$

Given that the sets of *leaders* and *candidates* are disjoint, maximizing the above likelihood function $\mathcal{L}(v)$ corresponds to maximizing the joint probability that the resulting respondent would nominate an individual from the current *candidates* other than an individual from the *leaders* set. Thus, following our proposed greedy approach for finding the minimal set of respondents, the individual corresponding to node $v_{\mathcal{L}} : \arg\max_v \mathcal{L}(v)$ is then surveyed, and the resulting nominations are added to the training set and fed back to the classifier.

As the previous approach relies solely on the classifier output for making the decision of who to survey next, the quality of the output is directly correlated with the accuracy of the underlying classifier. Therefore, a competing requirement for the proposed approach is to obtain labels for edges that will enhance the overall accuracy of the classifier.

In order to reduce the class probability estimation error, we follow the BOOTSTRAP-LV algorithm [14] which relies on estimating the local variance of each example in the pool, as an indicator of how well the model capture it given the available data. Instead of following the regular direct

Algorithm 1 Active Survey algorithm

Require: Input graph $G(V, E)$, classifier C, nomination threshold α, and required percentage of leaders k

Ensure: $\alpha > 1$

Set observed set $T_r \leftarrow$ Initial survey respondents
Set unobserved set $T_s \leftarrow V - T_r$
Set *leaders* $\leftarrow \{v : \exists e_1(u_1, v), \ldots, e_\alpha(u_\alpha, v) \in E \wedge class(e_1) = +, \ldots class(e_\alpha) = +\}$
Set *candidates* $\leftarrow \{v : \exists e(u, v) \in E \wedge class(e) = +\}$
while $\left(\frac{|leaders|}{|V|} \times 100 < k\right)$ **do**
  Set $v_B \leftarrow$ BOOTSTRAP-LV(C,T_r,T_s)
  Train(C, T_r)
  for each $v \in T_s$ **do**
    for each $e_x(v, u_x) \in E$ **do**
      $P(-|e_x(v, u_x)) \leftarrow$ Infer(C, $e_x(v, u_x)$)
    end for
    Compute log-likelihood $\mathcal{L}(v)$ according to equation 3.2
  end for
  Set $v_\mathcal{L} \leftarrow \arg\max_v \mathcal{L}(V)$
  Set $H_{avg} \leftarrow \frac{1}{n_{test}} \sum_{x=1}^{n} H(e_x)$
  With probability $p = H_{avg}$, set $v^* = v_B$, otherwise set $v^* = v_\mathcal{L}$
  Set $T_s \leftarrow T_s / \{v^*\}$
  Survey respondent v^*, update *leaders* and *candidates* sets according to the resulting nominations
  Set $T_r \leftarrow T_r \cup \{v^*\}$
end while

selection procedure of ranking the potential training examples by their effectiveness score, as most active learning algorithms do, the BOOTSTRAP-LV algorithm uses a weighted sampling approach to choose the next training example, which is shown to perform better than ordinary uncertainty sampling [14].

In order to provide a robust mechanism, we need to include both the objectives of maximizing the likelihood of obtaining a new opinion leader and minimizing the expected classification error in the decision of who to survey next. For that, we quantify the amount of uncertainty in the classifier output over all edges in the test set as:

$$H_{avg} = \frac{1}{n_{test}} \sum_{i=1}^{n} H(e_i(u, v))$$

where the entropy of the classifier output with respect to a given edge $e(u, v)$ is defined as:

$$H(e) = -\left[p(+|e(u, v))log(p(+|e(u, v))) + p(-|e(u, v))log(p(-|e(u, v)))\right].$$

Then, the next respondent to be surveyed v^* is chosen via a probabilistic decision based on the current accuracy of the underlying classifier as follows:

$$v^* = \begin{cases} v \sim BOOTSTRAP - LV & \text{with probability } p = H_{avg} \\ \arg\max_v \mathcal{L}(v) & \text{with probability } p = (1 - H_{avg}) \end{cases} \quad (2)$$

where the probability of choosing a respondent based on uncertainty sampling to enhance the classifier accuracy increases with higher uncertainty in the classifier output, while being more confident in the predictions yield higher probability of choosing a respondent to optimize the objective likelihood function. The full details of the model are presented in algorithm 1.

3.3 Initial Seeding

In most active learning methods, the choice of the initial training set is usually done at random or is assumed to be provided by an expert or an oracle. However, in our work we experiment with the effect of using different methods for choosing the initial seed for training in the proposed method.

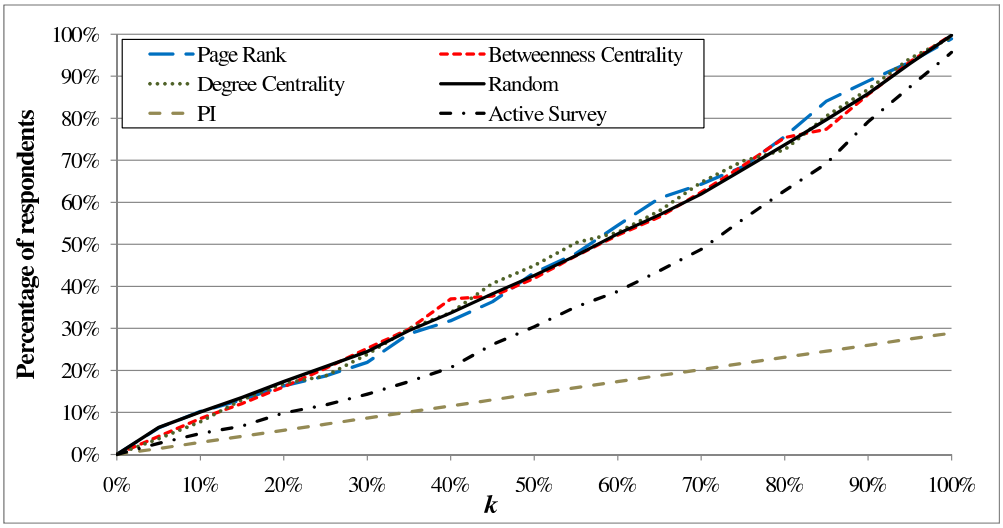

Figure 2: Percentage of respondents needed for various values of k at $\alpha = 2$

As our focus is geared towards identifying local opinion leaders within the physicians community, we compared a random method to a community-based one for choosing the initial seed. In our approach, we first ran a community detection algorithm on the primary geographical network, then choose a representative from each resulting community in the set of our initial survey respondents.

Our experiments showed that using the community based approach is indeed better than using a random method for choosing the initial survey respondents, with significant performance gains. Furthermore, we compared various methods for choosing the representative target from each community, and we found that choosing the individuals with the highest betweenness centrality from each community as a representative resulted in a slight increase on the performance of out approach, though not significant from other methods based on different centrality scores.

4 Experimental Evaluation

In order to test our proposed method, we use a health care dataset generously provided by *Community Analytics*[1], a leading research organization in social network analysis, which specializes in analyzing influence networks and identifying opinion leaders through conducting surveys with the target audiences of their clients. The data represents survey information for nominating locally physicians who influence the medical practice of the respondents.

The dataset consists of 2004 physicians, with 899 actual survey respondents generating 1598 nominations. As the surveys are based on identifying local influential physicians, we limit the potential edges for each respondent to the physicians whose locations are within 150 miles from her, yielding a set of 127,420 potential edges. By setting the nomination threshold (α = 2), we identify 260 opinion leaders in the network.

By using physicians' lists of publications from PubMed[2], we constructed both a citation and a co-authorship network among the physicians in the primary network. We also used the physicians affiliation information to construct a co-affiliation network as a third secondary source to leverage our primary data. Finally, using these three secondary networks, we generated the edge feature set on the primary physician network as well as enriched the node features with additional structural attributes from these networks.

To conduct our experiments, we use a logistic regression classifier and vary the target percentage k of opinion leaders to be identified, showing the corresponding percentage of respondents needed to

[1] http://www.communityanalytics.com
[2] http://www.ncbi.nlm.nih.gov/pubmed

obtain the target percentage by following our proposed active survey method. We compare our proposed approach with a random baseline and another method that uses a ranking of physicians based on their average centrality in the primary and secondary networks for determining the sequence of respondents to survey. We also show the performance of the perfect information (PI) method, which uses the actual labeled network (with full survey information) to choose the next respondent at each step. Note that the *PI* method represents the optimal solution at each point, and the lower bound for the number of required respondents for each value of k.

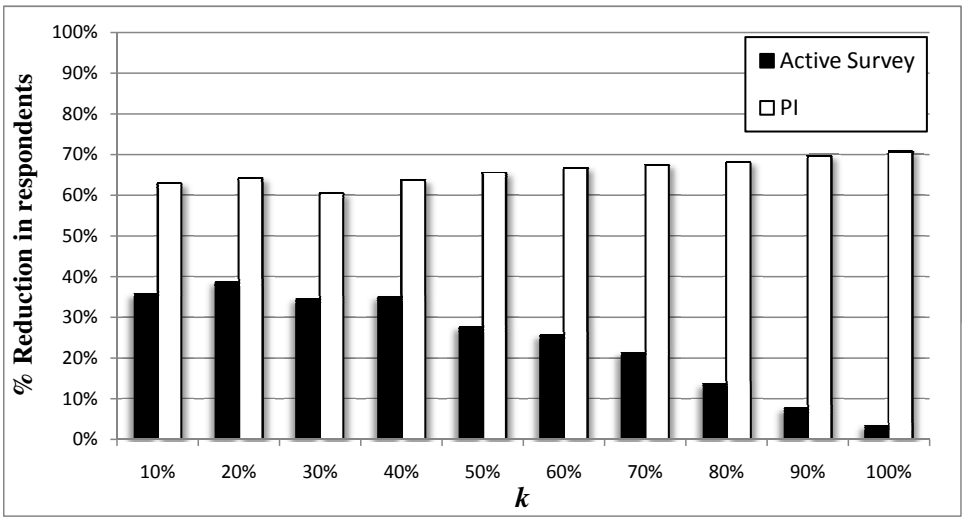

Figure 3: Reductions in required respondents to achieve various values of k at $\alpha = 2$

As can be noted in figure 2, although the performance of the methods based on various centrality measures is indistinguishable from the random baseline, our proposed active survey method's performance is significantly better than all other baselines. Figure 3 shows the actual percentage reduction in the size of respondent set of both the active survey method and the perfect information optimum with respect to the minimum set obtained by the best performing baselines at the corresponding value of k. As can be noted in the figure, our proposed approach yields around 35% average reduction in the number of respondents required. The reduction attained by the active survey method is reflected directly on surveying costs, thus helping the survey conductors to achieve their required goal with minimum cost.

5 Conclusion

In this work, we presented a novel active learning algorithm for prioritizing the acquisition of survey information for leadership identification. The approach requires the intelligent integration of both primary and secondary data, in order to identify which respondents to survey, based on both the likely information they will add to the candidate opinion leaders and also the utility of the information for improving the classification. We validated our results on a real-world dataset describing a physician nomination network.

Acknowledgments

This work was supported by Maryland Industrial Partnerships (MIPS) program under Grant #4409.

References

[1] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010.

[2] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[3] M. Gladwell. *The tipping point: how little things can make a big difference*. Back Bay Books, 2002.

[4] J. Turner. *Social Influence*. Open University Press, 1991.

[5] E. Keller and J. Berry. *One American in ten tells the other nine how to vote, where to eat, and what to buy. They are the influentials*. The Free Press, 2003.

[6] J. Myers and T. Robertson. Dimensions of opinion leadership. *Journal of Marketing Research*, (9):41 – 46, 1972.

[7] K. Chan and S. Misra. Characteristics of the opinion leader: A new dimension. *Journal of Advertising*, (19):53 – 60, 1990.

[8] D. Krackhardt. Structural leverage in marketing. *Networks in Marketing*, pages 50 – 59, 1996.

[9] T. Valente and R. Davis. Accelerating the diffusion of innovations using opinion leaders. *The ANNALS of the American Academy of Political and Social Science*, 566(1):55 – 67, 1999.

[10] S. Soumerai, T. McLaughlin, J. Gurwitz, E. Guadagnoli, P. Hauptman, C. Borbas, N. Morris, B. McLaughlin, X. Gao, D. Willison, R. Asinger, and F. Gobel. Effect of local medical opinion leaders on quality of care for acute myocardial infarction: A randomized controlled trial. *The Journal of the American Medical Association*, pages 1358 – 1363, 1998.

[11] Gaby Doumit, Melina Gattellari, Jeremy Grimshaw, and Mary Ann O'Brien. Local opinion leaders: effects on professional practice and health care outcomes. Cochrane Database of Systematic Reviews, 2007.

[12] D. Watts and P. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34(4):441 – 458, 2007.

[13] B. Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin - Madison, 2009.

[14] M. Saar-Tsechansky and F. Provost. Active sampling for class probability estimation and ranking. *Machine Learning*, 54(2):153 – 178, 2004.

[15] D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1994.