ABSTRACT
This paper predicts the stabilized tag set of a resource, with feedback of a small amount of user annotations, aiming to reduce the requirement of sufficient user annotations and to resolve the cold-start problem in a social annotation system.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; H.4.m [Information Systems]: Miscellaneous; H.5.3 [Group & Organizational Interfaces] Collaborative Computing

General Terms
Design, Experimentation, Performance

Keywords
Social Annotation, Spreading Activation, Tag Prediction

1. INTRODUCTION
With the progress of web 2.0, a variety of services promote web users from traditional information receivers to influential information propagators and even information sources. Social bookmark sites such as del.icio.us (http://del.icio.us) and Flickr (http://www.flickr.com) provide such web 2.0 services. Recently, applications and analyses of social annotations have attracted much more attentions [1][3][4]. We observe that limitations on the effectiveness of applying social annotations include: i) the requirement of sufficient annotations for a target resource; ii) the cold-start problem for resources newly arriving. To deal with the two issues, we design a supervised tag prediction model to predict the stabilized tag set for a target resource, with feedback of a small amount of user annotations.

Prediction of a stabilized tag set needs a complex model to capture how users make inference from a resource to its tags. Some researches [1][4] indicate three important characteristics in a social annotation system, i.e., i) keywords in a resource are usually selected as tags by annotators; ii) similar resources are usually annotated by similar tags; iii) shared background knowledge and imitative annotation between users would take influence on annotation activities. The basic idea of the proposed prediction model follows these characteristics. The following section describes the framework and the details of our model.

2. TAG PREDICTION MODEL
Our model predicts the stabilized tag set of a testing URL which has been annotated by several early users. The prediction algorithm is composed of two parts. In the first part (Section 2.1), we select candidate tags from the terms in the text content of a testing URL. We train a supervised scoring function from the training data to rank terms in the text. In the second part (Section 2.2), we combine the early annotations for the testing URL with the candidate tags, and then perform spreading activation [5][6] on a tag-correlation graph to find tags highly correlated to the combined tag set.

2.1 Content Tag Selection
Equation 1 is employed to score terms in a document (URL text content) for content tag selection. The basic idea of the scoring function follows the statistical translation model in information retrieval [2], which estimates the probability that a query would be generated as a translation of a document.

\[
CTSe(t_i \mid d) = \frac{\log(N \times \max(PC(t_i \mid term_j), \frac{1}{N}))}{\sum_{term_k \in \text{candidate}} \log(N \times \max(PC(t_i \mid term_k), \frac{1}{N}))}
\] (1)

In Equation 1, \(PC(t_i \mid term_j)\) denotes the probability of \(t_i\) as a stabilized tag, i.e., a tag included in the stabilized tag set, when \(t_i\) and \(term_j\) co-occur in the same document. \(N\) denotes the number of documents in the training data. \(PC(t_i \mid term_j)\) is estimated by Maximum Likelihood Estimation (MLE) over the training data. Assume \(D_j\) is the set of documents where \(term_j\) occurs and \(D_{ij}\) is the set of documents in \(D_j\) with \(t_i\) as a stabilized tag. Let \(PC(t_i \mid term_j) = |D_{ij}| / |D_j|\). If \(PC(t_i \mid term_j)\) is not 0, then \(PC(t_i \mid term_j)\) would be naturally as equal to or larger than \(1/N\). All terms in the document are ranked according to their scores. The content tag selection score of \(t_i\) will be normalized and then used as the initial activation score of \(t_i\) in the spreading activation of our prediction algorithm.

2.2 Spreading Activation
A tag-correlation graph is constructed for tag spreading activation. The tag-correlation graph is a network in which nodes represent tags and the edge(s) between two nodes represents their correlation. \(P_s(t_i \mid j)\), the probability of which stabilized tag sets containing \(t_i\) also contain \(t_j\), is an asymmetric correlation representing the strength of the edge from \(t_j\) to \(t_i\) in the directed graph. We also estimate \(P_s(t_i \mid j)\) by MLE over the training data.

The basic idea of spreading activation can be explained by a natural phenomenon. When we drop a stone into a pond, oscillation on surface transfers energy to neighborhood and becomes smaller and smaller in amplitude. Equation 2 and Equation 3 show how a node (i.e., a tag) \(t_i\) gains energy from its neighbors. At each iteration, node \(t_i\) propagates a portion \(\lambda_i\) of its energy to its neighbors, and gains some energy from its neighbors too. In our experiments, spreading activation is performed for a fixed number of iterations and eventually tags of the highest energies are proposed as the predicted stabilized tags.
Recall MAP annotations. performance, no matter whether 5 or 10 early annotations are testing sets, the stabilized tag set consists of the top 25 common by the highest content selection score of terms co-occurring with A tag is an with early user tags. Table 1. Performance of the content tag selection combined only 5 user annotations. More sophisticated approaches to estimate considerably large portion (~45%) of the stabilized tag set with a small amount of early user annotations. The experiment results show that our tag prediction model is able to predict a considerably large portion of each of the top 25 common tags of the highest frequencies.

3. EXPERIMENTAL SETUP
Our corpus is crawled from del.icio.us. In the corpus, total 15,934 URLs received more than 100 annotations form a sufficiently-annotated set. Total 2,000 URLs are randomly selected from this set as the testing data. Because the remaining URLs of the sufficiently-annotated set are not large enough for training, we increase the amount of training data by 45,156 more URLs annotated by at least 20 users. For each URL in the training and testing sets, the stabilized tag set consists of the top 25 common tags of the peak performances with different values of \( \lambda \) are very similar, while the curve of a larger \( \lambda \) shows more gradual. When \( \lambda \) is set to be 0.5, the performance in R-Precision of CTScr is improved from 0.3972 (see Table 1) to 0.4493 after 2 iterations of spreading activation. The relative improvement is 13.1% and is also significant. Recall and MAP are improved to 0.6072 and 0.3879, respectively. With spreading activation, the relative improvement in Recall is the largest (28.3%).

Figure 1. Performance of the tag prediction model with different parameters in the spreading activation

5. CONCLUSION AND FUTURE WORK
In this paper, we propose a prediction model for social annotation, aiming to predict the stabilized tag set of a web resource with a small amount of early user annotations. The experiment results show that our tag prediction model is able to predict a considerably large portion (~45%) of the stabilized tag set with only 5 user annotations. More sophisticated approaches to estimate scoring functions are necessary in the future work.

6. ACKNOWLEDGEMENT
Research of this paper was partially supported by Excellent Research Projects of National Taiwan University (97R0062-AE00-02) and Microsoft Research Asia (FY08-RES-THEME-087).

7. REFERENCES