To Tag or Not to Tag – Harvesting Adjacent Metadata in Large-Scale Tagging Systems *

Adriana Budura ⋆ , Sebastian Michel ⋆ , Philippe Cudré-Mauroux ◇ , Karl Aberer ⋆
⋆ EPFL, Switzerland
firstname.lastname@epfl.ch
◇ MIT, USA
pcm@csail.mit.edu

ABSTRACT

We present HAMLET, a suite of principles, scoring models and algorithms to automatically propagate metadata along edges in a document neighborhood. As a showcase scenario we consider tag prediction in community-based Web 2.0 tagging applications. Experiments using real-world data demonstrate the viability of our approach in large-scale environments where tags are scarce. To the best of our knowledge, HAMLET is the first system to promote an efficient and precise reuse of shared metadata in highly dynamic, large-scale Web 2.0 tagging systems.

Categories and Subject Descriptors H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval Models*; H.m [**Information Systems**]: Miscellaneous
General Terms Algorithms, Design
Keywords tagging systems, social communities, tag propagation

1. INTRODUCTION

The past few years have witnessed the rise of Web 2.0 applications promoting the sharing of documents through online communities. One standard practice is to rely on user-provided metadata to foster search capabilities. *Tags* are commonly used as user-generated metadata in Web 2.0 applications. Tags are essential in resolving user queries targeting shared documents, yet require human attention to be generated. Users typically tag a small fraction of the documents only, leaving most of the other documents with incomplete metadata.

We present a novel tag inference technique to assign tags to previously un-tagged documents or to extend the set of tags for already tagged documents by *propagating* existing tags from one document to similar documents. We put forward HAMLET (*Harvesting Adjacent Metadata in Large-ScalE Tagging Systems)*, a suite of principles, scoring models and algorithms for metadata propagation.

Our setting differs from traditional settings in two ways: (1) it is characterized by scarce information (due to the fact that the tagging process requires human input which is generally considered a scarce resource) (2) it relies on a user-defined organization of documents into a graph that connects related documents. For one particular document (i.e., the initial document), we gather the k most promising tags (according to our scoring model) from the document's neighborhood by propagating them along the edges of the document graph.

∗ The work presented in this paper was partially supported by the National Competence Center in Research on Mobile Information and Communication Systems (NCCR-MICS), a center supported by the Swiss National Science Foundation under grant number 5005-67322.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

2. PRINCIPLES & SCORING

In the following we present the principles underpinning our tag inference mechanism and show how they can be applied to create a scoring model to assess the relevance of a tag to a given document.

Indicators for Tag Relevance

∘ **Tag occurrence** is the frequency with which a tag appears in a given document neighborhood. It is meant to capture the popularity of a tag and it is expressed as a sum over all relevance scores of the given tag in the neighborhood.

∘ **Tag co-occurrence** is the frequency with which a pair of tags appear simultaneously. It takes into consideration the correlation among tags and is expressed in terms of an asymmetric measure of co-occurrence. In our work, we consider the conditional probability which is defined as $P(t_1|t_2) = P(t_1, t_2)/P(t_2)$, for two tags t_1 and t_2. As documents typically contain more than one tag we aggregate this measure to capture the relevance of a given tag t to a set of tags T by $\sum_{t' \in T} P(t|t')$.

∘ **Tag distance** is the distance between a document d' tagged with a tag t and a document d in the graph. Our assumption is that documents which are closer to each other in the graph have more related tags. The distance is expressed as the length of the shortest path between the documents d and d'.

∘ **Document similarity** is the similarity between the documents in our graph. We use the standard cosine similarity metric based on *relative term frequency* (rtf) values but our approach is not constrained by this particular choice.

Combined Scoring Function

Given an initial document d, we combine the four measures described above to assess the degree of relevance of a tag t observed in the neighborhood $N(d)$.

$$score(t, d) = \sum_{d' \in N(d) \wedge t \mapsto d'}^{m} S * O$$

The final score of tag t observed up to m times in the neighborhood of the initial document d is calculated by aggregating the partial scores for each of the single occurrences. The *sum* reflects our principle of *Tag occurrence* that favors those tags that frequently occur in the neighborhood. The parameter m controls the number of tag occurrences to be considered in the final aggregation.

The first factor, denoted as S, considers the relatedness of a document d' (for which a tag t is observed) to the initial document d:

$$S = \frac{\Pi_{i=0}^{|path(d,d')|-1} sim(d_i, d_{i+1})}{log(1 + dist(d, d'))}$$

This measure incorporates the *Document similarity* measure $sim(\cdot,\cdot)$ and the *Tag distance* $dist(\cdot,\cdot)$ as introduced before.

The second factor, denoted as O, expresses the relatedness of a tag t to the initial bag of tags of document d, which reflects the principle of *Tag co-occurrence* as mentioned above: $O = log(1 + co\_occ(t, T(d)))$.

3. TOP-K TAG SELECTION

In order to find the most relevant tags for an initial document, one solution would be to crawl the entire graph and rank all the tags according to their relevance to the initial document. This is in particular problemantic, if not impossible, in large scale distributed environments, or in cases where information from multiple portals have to be combined, for instance considering *Flickr* (http://flickr.com) pictures annotated via *del.icio.us* (http://del.icio.us). As a consequence, we consider tag inference by applying a top-k tag selection algorithm that carefully traverses the citation graph to minimize the number of documents visited.

In our model each node of the document graph represents one index list, consisting of $(tag, score)$-pairs where the scores are computed w.r.t. the initial document. Each edge in the graph leads from one index list to another. The algorithm prefers documents that have a higher probability to deliver promising tags. At any time during the traversal, the algorithm maintains a list of candidate tags, a list of the current top-k results, and a ranked list of documents to visit. Following the standard concepts of threshold algorithms [3], we continuously update the lower bound and upper bound scores of each seen tag and dismiss a tag if its upper bound score is smaller than the current rank-k result. The lower bound score considers those scores that have already been reported for the tag whereas the upper bound score denotes the lower bound score plus the largest possible scores coming from unexplored regions of the graph. A tag observed m' times so far can gain up to $(m - m') * s_{max}$ score mass, where s_{max} is the maximum score a tag can get if observed at the currently most promising document to come. The algorithm stops when the candidate list is empty and $m * s_{max}$ is smaller than the rank-k score.

4. EXPERIMENTS

We confirm the validity of our approach by using documents and their attached tags from *CiteULike* (http://citeulike.org), a popular tagging portal for academic papers. We construct the document graph by adding an edge $d_i \rightarrow d_j$ whenever d_i cites d_j. The citation graph has been obtained from *CiteSeer* (http://citeseer.ist.psu.edu).

Our dataset consists of $\sim 540K$ papers and $\sim 195K$ distinct tags. For our performance evaluation we randomly selected 10 initial papers having on average 12 citations. We gathered for each of the 10 papers all tags occurring up to level three in the document graph. Finally, we asked twenty colleagues to evaluate the relevance of each tag w.r.t. its corresponding initial paper.

We report on the precision@k values, i.e., the fraction of relevant tags among the top-k returned tags, and also on the number of nodes visited in the document graph. For the precision measure, we disregard all inferred tags that were among the initial tags of the document. Note that we do not report on recall since (i) we focus on top-k retrieval and (ii) the number of relevant tags for a document is potentially unlimited.

m	k=3	k=5	k=7	k=10	k=15	k=20	k=25
3	0.77	0.70	0.64	0.63	0.59	0.52	0.50
5	0.77	0.68	0.64	0.63	0.61	0.52	0.50
7	0.77	0.66	0.64	0.64	0.61	0.52	0.50
10	0.77	0.66	0.66	0.63	0.61	0.52	0.50
15	0.77	0.70	0.66	0.63	0.61	0.54	0.50
20	0.77	0.70	0.66	0.63	0.61	0.54	0.50

Table 1: Precision@k

m	k=3	k=5	k=7	k=10	k=15	k=20	k=25
3	59	71	69	87	106	106	137
5	71	91	103	119	136	130	182
7	81	110	110	138	153	160	194
10	99	123	139	157	177	170	199
15	126	132	163	172	188	196	202
20	134	149	170	182	193	199	204

Table 2: Number of visited documents

Table 1 reports on the precision@k for changing values of m. Our precision@k values are very promising in particular for $k < 10$, where we achieve 77% for $k = 3$ and close to 70% for $k = 5$. Not surprisingly, the precision decreases for larger values of k. Table 2 reports on the number of visited neighbors. With increasing m more neighbors are visited since m directly influences the upper bound scores of observed tags which are essential for the candidate pruning and hence the stopping of the algorithm. Another observation is that the influence of m on the precision@k is almost negligible which supports our initial hypothesis that the most promising tags are actually found in the close vicinity of a paper and that exploring further regions does not contribute to the top-k result.

5. RELATED WORK & CONCLUSION

Recently the sphere of social annotations has gained increasing attention. [5, 6, 4] have concentrated on understanding the tagging process and the resulting social annotations. [1] introduces two algorithms which integrate information extracted from the tags into the search process. These approaches are actually orthogonal to our efforts since they do not address the problem of inferring tags for resources. [2] generates personal annotations but focuses solely on the content of the documents.

We have put forward HAMLET, a suite of principles, scoring models and algorithms for metadata propagation. Our methods are based on a socially-driven graph relating similar documents. The techniques we have developed can obviously be extended beyond our showcase scenario focusing on academic papers and citations. We believe that the approach proposed in this paper could constitute a pivotal cornerstone towards improving information retrieval in large-scale annotation systems.

6. REFERENCES

[1] S. Bao et al. Optimizing web search using social annotations. *WWW*, 2007.

[2] P. A. Chirita et al. P-tag: large scale automatic generation of personalized annotation tags for the web. *WWW*, 2007.

[3] R. Fagin et al. Optimal aggregation algorithms for middleware. *J. Comput. Syst. Sci.*, 66(4), 2003.

[4] S. Golder et al. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2), 2006.

[5] H. Halpin et al. The complex dynamics of collaborative tagging. *WWW*, 2007.

[6] C. Marlow et al. Ht06, tagging paper, taxonomy, flickr, academic article, to read. *HYPERTEXT*, 2006.