ABSTRACT
This study investigated the retrieval effectiveness of tagging through experimental tests. We developed a test collection, a list of topics and corresponding relevance judgments using data collected from Pubmed and CiteULike. Two retrieval test runs were carried out, one without tags (baseline) and one with tags (tag run). General improvements in retrieval performance were observed from the tag run in terms of mean average precision and the difference is statistically significant though small. As revealed by the 11-point precision recall chart, the tag run retrieved more relevant documents at the top of the result list overall. The initial findings suggest that the use of tags in information retrieval has promise, especially in ranking the top retrieved documents.

Keywords
tagging, information retrieval, retrieval effectiveness, experimental study.

INTRODUCTION
Tagging has attracted wide attention. According to PEW research center surveys (Rainie, 2007) 28% of internet users have already indexed online content with tags and 7% do so several times a day. Saur (2009) discusses the potential impact of tagging and folksonomies in indexing and information retrieval. Folksonomies employ user generated tags applied by users instead of controlled vocabulary keywords assigned by a trained professional lowering the cost of producing metadata. However, the ability of tagging to enhance traditional methods of indexing will depend on how effective tags are at improving retrieval performance. The purpose of this study is to investigate whether tags can improve information retrieval and how effective they are at enhancing retrieval performance. Empirical answers to these questions will help support decisions on the use of tagging for retrieval.

EXPERIMENTAL DESIGN
The experimental design of this study follows TREC conventions (http://trec.nist.gov/) including the creation of a test collection, a list of topics and relevance judgments. Common evaluation measurements including average precision, recall and 11-point precision recall charts are used to compare runs. We employed the Indri search engine as the test retrieval system (Strohman et al., 2004).

Test Collection
We collected 234,338 records in CiteULike with a Pubmed ID. Tracing these Pubmed IDs, we collected the corresponding Pubmed record from Pubmed, which included the article title, MeSH headings and abstract. Then, the article titles, MeSH headings and abstracts were combined with tagging data from CiteULike according to their Pubmed IDs. This produced a set of Pubmed articles with their corresponding tags from CiteULike. To conform to standard indexing conventions, tags consisting of multiple terms concatenated by a hyphen or underline were separated. Using the Indri search engine, the article title, abstract, Pubmed ID and user-generated tags were indexed for retrieval.

Topic Selection
Similarly to TREC-9 (Voorhees & Harman, 2000), we used the MeSH headings as topics. Since each article in Pubmed was assigned MeSH headings by professional indexers, the article is considered to be a relevant result for each of its headings. Fifty topics were randomly selected from all collected MeSH headings. Queries were manually constructed from the terms in the MeSH heading, Scope Note, Previous Indexing and Entry Terms fields of the relevant MeSH record. Each author generated their own queries independently. Overlapping terms were used for the experiments. Domain expert suggestions were employed for controversial terms. The average query length is 5.5 terms. The shortest query contains 2 terms and the longest contains 18 terms. Query terms were submitted to Indri search engine, which adopts inference network to combine the belief of different terms.

Controlled Test
We conducted two test runs to compare retrieval performance with tags and without tags. In the baseline run, only the article title and abstract were searched using the fifty topics
we generated. In the tag run, article title, abstract and tags were searched with the same queries. Comparing these two runs will reveal how tags affect retrieval performance.

RESULTS
General improvement was observed from the tag run compared to the baseline in terms of mean average precision (MAP), namely 0.231 for baseline and 0.242 for the tag run. There is a 4.76% increase overall and the difference is statistically significant at the 0.05 level as indicated by the paired Wilcoxon test (p=0.016, N=50).

![Figure 1. Difference between tag run and baseline for each topic in terms of average precision](image1.png)

Figure 1 illustrates the effect of tags, showing the difference between the two runs (tag run minus baseline) in terms of average precision. The effect size varies by topic, which indicates further study is needed to understand how tags work. Besides MAP, we also calculated average recall and average precision values at the last relevant retrieval for 50 topics (Table 1). Overall, a 3.52% improvement on average recall was observed which indicates that the tag run retrieved more relevant documents than the baseline. A slight decline (-0.01%) in average precision at the last relevant retrieval was found, which means the baseline run had a very weak advantage on precision towards the end of the result list.

![Figure 2. Overall 11-point precision recall chart](image2.png)

The tag run outperformed the baseline at the beginning of the retrieval results overall. From the recall point of 0.1 to the end, the difference between the runs is limited. Therefore, tags improve retrieval performance by retrieving more relevant documents at the top of the result list. This finding is important to information retrieval since previous studies (Spink et al., 2001) have revealed that users seldom check results that are ranked beyond the third page.

CONCLUSION
The initial results indicated that tags improved the overall retrieval performance for the 50 topics we constructed though the increase is small. The variations in effect size suggest further study is needed. Additionally, the tag run was able to retrieve more relevant results and rank them at the top of the result list. These findings suggest that tags are potentially promising for retrieval. Future study will extend our experiments to a full text environment to further examine the effectiveness of tags.

REFERENCES


<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Tag</th>
<th>Improvement/decline against baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.231</td>
<td>0.242</td>
<td>4.76%</td>
</tr>
<tr>
<td>Average Recall</td>
<td>0.454</td>
<td>0.47</td>
<td>3.52%</td>
</tr>
<tr>
<td>AP@last relevant retrieval</td>
<td>0.2757</td>
<td>0.2756</td>
<td>-0.01%</td>
</tr>
</tbody>
</table>

Table 1. Comparison between baseline and tag run

To compare the precision values at different recall levels, an overall 11-point precision recall chart was plotted (see Figure 2).