Indexing Quality and Effectiveness: An Exploratory Analysis of Electronic Theses and Dissertations Representation

Daniel Gelaw Alemneh and Mark Edward Phillips University of North Texas, USA

INTRODUCTION

Theses and dissertations (ETDs) represent a wealth of scholarly and artistic content created by graduate students in masters and doctoral programs in the degree-seeking process. Considering the multi-disciplinarity and interdisciplinarity characteristics of ETDs, often several subjects and indexing terms need to be supplied to adequately represent ETDs for efficient access.

INDEXING AND INFORMATION RETRIEVAL

An index is simply a systematic representation of an information-bearing object (text, images, or other resource) that helps users find needles in the information haystack. It points users to specific items on topics of interest. It shows users related topics and indicates information trails through vast information stores. To fully understand what an index is, it is necessary to be both micro- and macro-minded. On the micro level, we concern ourselves with the specific mechanisms of creating an index. On the macro level we put an index into the larger context of an information retrieval system. Figure 2 shows the inverse relationship of precision and recall.

A good index helps users find what they need, even when they are not sure themselves what they need. Traditionally, inter-indexer consistency is a measure of indexing quality or effectiveness. However, as depicted in Figure 3, digital contents exhibit many aspects that make representation difficult. Recognizing the fact that abstruseness is often in the eye of the beholder, the subjectivity and objectivity of the process and the need to distinguish functional representation from mere descriptions of a topic have been emphasized.

METHODOLOGY

This study analyzed the index terms in UNT's ETD Collection from two sides -- the document side (supplied by creators and librarians) and the user side. It specifically tried to gauge the effectiveness of the terms in matching users' queries. At the time this research was conducted, 11,873 unique items were available with metadata. To get a better sense of users' discovery of digital resources, we also looked at the following two questions:

1. Were users arriving at our digital resources from searches that were answered by an item's descriptive metadata or by parts of the full text of the item?
2. If or when a resource was found with metadata, which fields were being used to retrieve that item?

METHODS

Table 1 and 2 show statistics for the number of queries per unique item and record found based on matches in metadata and full text (occurred between May 4 2014 and January 24, 2016). Although queries varied in length, they were analyzed as individual words (tokens) rather than phrases. This allowed for partial matches in a given field.

<table>
<thead>
<tr>
<th>N</th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
<th>Sum</th>
<th>Mean</th>
<th>Stddev</th>
</tr>
</thead>
<tbody>
<tr>
<td>43420</td>
<td>1</td>
<td>2.5</td>
<td>31</td>
<td>104102</td>
<td>2.40</td>
<td>1.59</td>
</tr>
</tbody>
</table>

Table 1. Statistics for the number of tokens per query.

Figure 4 shows how many items could be found using either index equally, how many had a partial match in one index with a full match in the other index, and the number of queries that could be found only through the combination of metadata and full text versus either index alone.

Considering the interdisciplinary nature of ETDs and the diverse global user communities, effective retrieval depends not only on the indexing terms assigned to describe ETDs, but on the search query terms entered by users. Moreover, effective metadata and taxonomies add value and amplify the (mostly interdisciplinary) ETDs–allowing users to explore and delve deeper in multidimensional ways.

REFERENCES


Figure 1. The Usage Statistics for UNT ETDs, as of October 10, 2016: http://digital.library.unt.edu/explorer/collections/UNTETDs.html

Figure 2. Depicting Relationship of Precision and Recall

Figure 3. Depicting Relationship of Precision and Recall

Figure 5. Five Basic Aspects of a Digital Object (Modified from van Wijngaarden, 2007)

Figure 4. Record discoveries categorized for dual metadata and full text usage.

Figure 5. Record discoveries categorized for dual metadata and full text usage.