Information Retrieval from Digital Libraries: Assessing the Potential Utility of Thesauri in Supporting Users' Search Behaviour in an Interdisciplinary Domain

Ali Shiri

Abstract. The objective of this research was to investigate the extent to which thesauri have the potential to support the search behaviour of nanoscience and technology researchers while interacting with an electronic book digital library. Transaction log data was obtained from a nanoscience and technology digital library to investigate the nature, type and characteristics of users' queries and search terms. The specific objectives was to assess the extent to which users' search terms matched with those found in two well-established thesauri attached o the INSPEC and Compendex databases.

1. Introduction

Nanoscale science and technology have seen rapid growth and expansion in new areas in recent years. Due to the interdisciplinary nature of this area, nano researchers need to consult a variety of resources. Kutner (2000) examines challenges and issues in searching in interdisciplinary areas. She argues that although electronic bibliographic databases that provide sophisticated searching capabilities and multiple access points to the scholarly literature have been a boon to the interdisciplinary researcher, problems continue to exist in terms of lack of consistent interfaces and consistent controlled vocabularies across databases. Although there are a number of studies focusing on search behaviour of interdisciplinary researchers, there is little research on whether thesauri have the potential to support interdisciplinary search behaviours. There is a growing interest in reusing and repackaging thesauri and other types of controlled vocabularies in various digital libraries, institutional repositories and content management systems. In order to assess the utility and reusability of thesauri in new information environments, research needs to explore the type, nature and characteristics of user terms and queries within the context of new information repositories such as digital libraries. This type of research will shed light on the ways in which thesauri can be reused and/or redesigned to accommodate users' searching needs. This study has investigated users' queries and search terms as revealed by transaction log analysis of a nanoscience and technology digital library. Moreover, it has assessed the extent to which the search terms have exact or partial match equivalents in two thesauri attached to the INSPEC and Compendex databases.

2. Methodology

The theoretical framework upon which this study was carried out can be found in a number of user-thesaurus interaction studies (Shiri & Revie, 2004, 2006).

2.1 Research Questions:

The following research questions were formulated to address the objectives of this study.

- What types of queries do NanoNetBase users formulate? Subject searches vs. known- item searches?
- What are the characteristics of the queries in terms of the number of terms?

- How many of the user terms did match those found in the INSPEC and Compendex Thesauri?

- How many percent of search terms are Exact matches, how many percent are partial matches?

2.2 The System: NanoNetBase E-book Digital Library

NANOnetBASE is an e-book digital library for nanotechnology and nanoscience researchers. The full-text database consists of 45 titles that can be accessed in a variety of ways. The library provides a variety of search tools, such as Boolean operators, truncation, wildcard, stemming, fuzzy searches, field searching, phonic search, synonym search, and variable term weighting.

2.3 Transaction Log Dataset

Transaction log analysis of web search engines, intranets, and websites can provide insight into understanding of the information searching process of online searchers (Jansen, 2006). The data used in this study consisted of transaction logs from the NANOnetBASE, a digital library subscribed and used by a Canadian university. Transaction logs from July 2004 to October 2006 were examined. In total, 1921 transactions were analyzed. The transaction logs contained the following information: date and time of activity, user identifier (in the form of IP address), activity detail (query), activity type: book viewed, advanced search, quick search, search terms, search results viewed, or managed account information.

2.4 Thesauri Used

Two thesauri were utilized for this study, namely INSPEC and Compendex. These thesauri are part of the INSPEC and Compendex databases. The reasons for the choice of these thesauri are a) they are both well-established thesauri and b) the cover many aspects of nanoscience and technology. The users' search

terms derived from transaction log analysis of NanoNetBase were all compared against these two thesauri.

2.5 Data Analysis

Three levels of analysis were defined, namely term-level analysis, query-level analysis and session-level analysis. For the purpose of this paper we will particularly focus on the first two levels. A list of search terms and queries was gleaned from the transaction log data set to allow for the analysis. All of the terms were compared with those in the INSPEC and Compendex thesauri.

3. Results

The results below are arranged based on the order of the research questions.

Research question1: *What types of queries do NanoNetBase users formulate? Subject searches vs. known- item searches?*

Query types	**No of queries**	**Percentage of queries**
Subject queries	333	84.74%
Title queries	32	8.14%
Author queries	3	0.76
Unknown	25	6.36
Total	393	100

Table 1.Types of user queries

As Table 1 shows, subject queries constitute around 85% of the queries submitted by users. This is particularly important as it shows that subject queries are very popular within the context of digital libraries.

Research question 2: What are the characteristics of the queries in terms of the number of terms?

Table 2 shows the number of queries along with the number of search terms in The average number of terms used was 2.11 and the largest number of terms used was 8.

Number of terms used in query	**Number of Queries**	**Percentage of Queries**
1	156	40%
2	137	35%
3	53	14%
4	19	5%
5	9	2%
6	13	3%
7	3	n/a
8	3	n/a

Table 2. Number of query terms in searches

Research question 3: How many of the user terms did match those found in the INSPEC and Compendex Thesauri?

Research question 4: How many percent of search terms are Exact matches, how many percent are partial matches?

In order to assess the extent to which users' terms mapped to the thesaurus, all terms entered by users were analyzed and their mapping situations are reported here. Table 3 shows the details of the user terms matched the INSPEC thesaurus along with the number of terms for each match type.

Match type	Number of terms	%
Exact match	152	37
Partial match	51	12
No match	174	43
Broader term match	8	2
Narrower term match	23	6
Total	408	100

Table 3. Number and percentage of search terms entered by the user and matched in various ways to the INSPEC thesaurus

The analysis in Table 3 shows that around 50% of the terms entered by users are matched those of the INSPEC thesaurus.

Match type	Number	%
Exact match	199	49
Partial match	53	13
No match	135	33
Broader term match	7	2
Narrower term match	12	3
Total	406	100

Table 3. Number and percentage of search terms entered by the user and matched in various ways to the Compendex thesaurus

Table 4 shows that around 62% of users' search terms were matched those of the Compendex thesaurus.

These findings suggest that these two thesauri have the potential to assist users in their search process. The results of this study imply that these thesauri can be

reused in such a digital library as NanoNetBase. From an information retrieval perspective, the INSPEC and Compendex thesauri have the potential to support both interactive and automatic query formulation and expansion processes. This research makes a number of contributions to the areas of knowledge organization, digital libraries and interdisciplinary search behaviour studies. It will offer some Implications for the design of thesauri and controlled vocabularies for interdisciplinary research areas. In particular, it identifies the areas that these thesauri lack in terms of subject coverage and specificity. Another contribution of this research is that it demonstrates the coverage of the INSPEC and Compendex databases and their associated thesauri in relation to nanoscience and technology. Since nanotechnology is a relatively new and multi-disciplinary field, this study will provide insight into interdisciplinary search beahviour and how users search and how the information access and retrieval interfaces may better be constructed so that users can access the information they need in an efficient and effective way.

References

Jansen, B. J. (2006) Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research*, 28(3), pp. 407-432.

Kutner, L.A. (2000) Library instruction in an interdisciplinary environmental studies program: challenges, opportunities, and reflections. *Issues in Science and Technology Librarianship*, No. 28, available from: http://www.library.ucsb.edu/istl/00-fall/.

Shiri, A. A.; Revie, C. (2004) End-user Interaction with Thesauri: An Evaluation of Cognitive Overlap in Search Term Selection. *Proceedings of the 8th International Conference of the International Society for Knowledge Organization (ISKO)*, London, 13-16 July 2004

Shiri, A.; Revie, Crawford (2005) Usability and User Perceptions of a Thesaurus-enhanced Search Interface. *Journal of Documentation*, 61(5), 640-656. (Received the Journal of Documentation 2006 Highly Commended Award)