Paper: EH

Thesaurus in an Automated Information Retrieval System

Swapan Manna

Documentation Research & Training Centre

Indian Statistical Institute

Bangalore 560 059

email: *m_swapan@yahoo.co.in*

Extended Abstract

The world of information retrieval today has changed dramatically, with immense increase in the availability of searchable full text and the increasing availability of powerful search engines. Still search engines retrieve more junk than pinpointed information necessitating the use of more efficient retrieval tools. Hence it is reasonable to ask whether there is any place left for thesauri in the new information retrieval environment. Meanwhile the use of machine-aided or even automatic indexing has also been raising a demand for the use of thesauri. A thesaurus is a tool designed to aid users in finding their way around the vocabulary of a database. In addition to its traditional use as an authority for the terms used in indexing the database it offers reminders to terms the user might not even have considered. In the modern context a thesaurus could be extremely useful to provide controlled access to large collections of text and unstructured information, and to help search engines to have more

precision. Intelligent retrieval systems, which integrate statistical and semantic information to retrieve more useful results, could make use of an extensive thesaurus of word types and relationships. Producers who are concerned with providing standardized subject access to their resources can also make use of thesauri to determine the content of the element(s) allocated to subject metadata. Certain fundamental problems are observed in the basic design of thesauri that make them less than optimally useful for more powerful retrieval scenarios. Still there are both pragmatic and philosophical reasons to be positive about the need for continued use of thesauri. And it is hoped that controlled vocabularies will be the foundation of next-generation web sites and Intranets.