INTERNATIONAL JOURNAL OF INFORMATION SYSTEMS

ISSN:2229-5429

(A Journal of SIMCA)

Vol IV, Issue II, Jan-May 2014

RESEARCH JOURNAL

Indexed By :

J-Gate

Largest E-Journals Gateway

Unstructured Data Mining - Methods and Techniques for Information Retrieval

Prof. Leena Deshmukh, AP
JSPM's Jayawant Institute of Management Studies, Tathawade, Pune
linadeshmukh@gmail.com
9960639644

ABSTRACT

Textual data is unstructured. The term information retrieval generally refers to the querying of unstructured textual data. Unstructured data mining adopts different techniques to discover and retrieve information from large collection of documents. Data contained in documents are unstructured without any associated schema. The process of information retrieved consists of locating relevant documents. In this paper commonly used methods and techniques like Text Retrieval - the query is regarded as specifying constraints for selecting relevant documents (Document selection, Document ranking, Vector space model), Text Indexing (Inverted indices, Signature file), and Query Processing (Relevance feedback – when examples of relevant documents are available system can learn from such examples to improve retrieval performance, pseudo-feedback – when we don't have such examples, a system can assume a top few retrieved documents tobe relevant & extract more related keywords), for retrieval of unstructured data are discussed with their merits and demerits.

Keywords: Data mining, tools, precision, recall, F-score

INTRODUCTION

Now a day's most knowledge seekers are searching data through internet. This information is stored in the text databases or document databases, which consists of large collection of documents from various sources, such as news articles, research papers, books, digital libraries, e-mail messages and web pages. Text databases are rapidly growing due to increasing amount of information in each sector.

Data stored in text databases are semistructured data in that they are neither completely unstructured nor completely structured. There have been a great deal of studies on the modeling and implementation of unstructured data in recent database research. Information retrieval techniques, such as text indexing methods, have been developed to handle unstructured documents.

Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data. Typically, only a small fraction of the many available documents will be relevant to a given individual user. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the

49