ABSTRACT

Information retrieval includes locating unstructured text material documents which satisfies information need from large collections. Information retrieval tries to locate documents in a collection about a given topic or which satisfies a specific information need. The topic or information need is expressed through a user generated query. According to users documents which satisfy a query are relevant and which do not about a topic are non-relevant. A query may be used by an information retrieval engine to classify a document collection (or in an incoming stream), returning a document subset which meets some classification criterion for the user. The higher the proportion of user returned documents that are relevant, the better the classification criterion.

A basic information retrieval concept is measuring similarity between two documents. For this, a small word set input into a search engine is considered, a document to be matched to others. Measuring similarity is related to predictive methods for learning the classification methods. Measuring similarity is a common theme and the method’s variations are basic to information retrieval.

The advent of internet technology to share educational contents and experiences ensured that institutions globally offered a federated search to courses, lesson plans, contents, assignments, seminars and experiments, all of which are stored in repositories of learning content management systems
controlled by a learning management system. The problem faced by a learner’s community is in accessing, sharing and delivering quality relevant content for online teaching learning systems. Today peer to peer networks are used for daily sharing of videos, audios, images, music or other distributed learning digital processes. Hence, sophisticated search and information retrieval solutions are necessary. Many existing web based course structures developed in e-learning system are considered as course ontology and can be mapped into a model. This ontology based solution increases information retrieval accuracy through high precision and recall.

This study addresses pre-processing and various source documents retrieval to achieve improved Information Retrieval systems and investigates tools/techniques used for autonomous classification or documents clustering; new methods are proposed based on concept expansion. Reuter’s dataset and Movielens dataset are used in this research. The following summarizes the objective of the thesis:

i. Propose a concept expansion for creating corpus

ii. Propose a Language modelling based on Nominal Language Model

iii. Propose a cluster based Feature selection method based on Particle Swarm Optimization and Genetic Algorithm

iv. Optimization of document clustering using proposed Mobile Agent
A term or phrase has multiple meanings, though a domain specific concept is unambiguous. It is useful to use documents domain specific concepts rather than retrieving documents terms from a specific domain. Nominal language model based approach includes natural language processing lexical resources where the process goes through data extraction with a query. Experiments showed that proposed dynamic nominal language model method achieved better recall compared to traditional tf.idf and language modelling using query likelihood. A hybrid optimization based feature selection method was proposed to solve nonlinear partitioned clustering problem. The hybrid algorithm is based on particle swarm optimization and genetic algorithm. Experiments reveal that proposed feature with proposed dynamic nominal language model had better recall compared to traditional tf.idf and language modelling using query likelihood.

Two important issues with query expansion are addressed: selection and weighting of additional search terms. Compared to earlier methods, the proposed method expands queries by adding terms similar to query concept, rather than selecting terms similar to query terms. Experiments conducted for varying number of additional terms (5, 10, 15, 20, 25) and results proved superiority of the proposed concept based query expansion method with respect to precision. It is also seen that 15 additional terms achieve maximum precision. This study aims to demonstrate mobile agent technology’s ability in information retrieval. The system demonstrates workability of this technology combination in an e-Learning application. In mobile agent concept, 75 additional words are generated and 5 queries used by each mobile agent.
Mobile agent technology has been used for distributed information retrieval. The trained or the intelligent mobile agent carries different reformulated query to different heterogeneous learning content management systems. It is shown that the precision values obtained for tdf. idf by improved 17.2% when recall is 0.01 and 25.1% when recall is 0.9. The proposed technique performed better than concept expansion by 9.6% when recall is 0.01 and 0.9. Similarly for Language modelling using query likelihood, the proposed technique performed better by 12.3% when recall is 0.01 and 16.4% when recall is 0.09.