With changing user expectations, many traditional libraries are moving toward digital content storage. Accessible from anywhere at any time, digital contents as stored in digital libraries provide users with efficient, on-demand information experiences. With this trend, the amount of digital contents especially digital text documents made available to users have tremendously increased over the years, being filled with hidden information in form of the varieties of topics of discourse inherent in them leading to information overload. Accordingly, users, mostly computational researchers are presented with challenges on the discovery and identification of the varieties of topical contents of the collections in the digital library thus making it imperative to develop a means to automatically discover the topics that pervade the collections in a digital library. This paper therefore presents UPH Digital Library Miner, a software application for mining document collections of a digital library for topical structure discovery and topic-based similarities search between collection pairs, using topic modeling algorithm and inverted Kullback-Leibler divergence measure. The application is integrated with document collections
UPH Digital Library Miner: A Topic Modelling-based Software Application for Mining Document Collections

built in a widely used digital library software system— Greenstone digital library system, via loose-coupling integration approach. Results obtained from using this software application on the Greenstone’s document collections that contain abstracts of about 628 documents from IEEE transactions on Software Engineering show its ability to discover latent topical structures in collections and also report collections that are similar based on their discovered topical structure.

References

Index Terms

Computer Science

Information Sciences

Keywords