Abstract

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

the Association of Computational Linguistics. College Park MD, Association of Computational
Linguistics, Morristown, NJ. pp.3-10.
27, No. 4, pp. 245-250.
Analyzing Unstructured Data, New York: Cambridge University Press.
International Conference on Research and Development in Information Retrieval, pp. 50-57.
Laurence Erlbaum.
8. Mimno D. and McCallum, A. 2007. Mining a Digital Library for Influential Authors. In
JCDL’07 ACM, Vancouver, British Columbia, Canada, June 18–23.
Self-Instructional Guide for Beginners. Thrissur, India.
Representation of Topics and Genres. Applied Intelligence, vol. 18, 271–293.
Overview. International Journal of Information Technology and Knowledge Management, vol. 2,
no. 2, pp. 613-622.
Framework for Mining Digital Library’s Text Documents. IEEE African Journal of Computing and
ICTs, Nigeria, vol. 8, no 4.
hmallet.cs.umass.edu.
Between Probabilistic Automata Using Rational Kernels. Harvard University, Division of
Engineering and Applied Sciences, Cambridge.
http://nlp.stanford.edu/software/tmt/tmt-0.4..ieeeexplore.ieee.org/xpl/tocresult.jsp?
isnumber=28304&punumber=32.
Index Terms

Computer Science

Information Sciences

Keywords