Literature digital libraries, now an indispensable part of research and education worldwide, are increasing in size at very high numbers. As an example, PubMed, a literature digital library for biomedical sciences, currently contains 15 million papers, and is increasing at a rate of 400,000 papers every year. This issue of Data Engineering Bulletin is on the critical areas of searching, mining, querying, and information extraction from literature digital libraries.

In "Scaling Information Extraction to Large Document Collections", Eugene Agichtein classifies and reviews four approaches for scalable information extraction from large document collections, namely, scanning large document collections, exploiting general-purpose search engines, employing specialized indexes and search engines, and using parallelization and distributed processing. Algorithmic approaches trade off information extraction accuracy and completeness for speed. A promising approach is to store semantically annotated documents in semi-structured form.

Text analysis engines are different than search engines in that they allow for queries with words and entities such as punctuation, tags, etc. as well as returning results of different types, e.g., sections and phrases of documents. In "Fast and Furious Text Mining", Joel D. Martin describes and briefly evaluates the performance of a text analysis engine called TLM ("Text and Language Mining") with a highly expressive query language. TLM is part of an integrated suite of tools called LitMiner.

Example-based publication searching is becoming common place in digital libraries, which essentially requires the evaluation of a publication similarity measure. In "Evaluating Publication Similarity Measures", Sulieman Bani-Ahmad, Ali Cakmak, Gultekin Ozsoyoglu and Abdullah Al-Hamdani classify the existing publication similarity measures as text-based (from Information Retrieval) and citation-based employing bibliographic coupling and/or co-citation, and extend and evaluate a number of publication similarity measures in terms of accuracy, separability, and independence.

Current search engines are known to perform poorly for a number of "hard" queries. In "Hard Queries can be Addressed with Query Splitting Plus Stepping Stones and Pathways", Xiaoyan Yu, Fernando Das-Neves, and Edward A. Fox propose an approach based on "Stepping Stones and Pathways" and query splitting, and find the approach feasible and promising.

In "Ten-Year Cross-Disciplinary Comparison of the Growth of Open Access and How it Increases Research Citation Impact", Chawki Hajjem, Stevan Harnad, Yves Gingras report that openly accessible (OA) articles from ten disciplines are cited more than those that are not. Their results indicate that the overall percentage of OA articles varies from 5% to 16%, and OA articles have from 25% to 250% more citations as compared to non-OA articles.

Finally, in "A System of User-Guided Biological Literature Search Engine", Meng Hu and Jiong Yang propose and briefly evaluate a new digital library search paradigm based on iterative clustering and user feedback.

I hope that you will find this issue useful and informative. My special thanks to all the authors for their contributions to this special issue of the Bulletin.