

56. Natural Language Processing and Information Retrieval Techniques Applied to the Reproductive Biology Literature

R. Farhan, J. D. Aplin, T. K. Attwood, M. McGee Wood and C. P. Sibley
University of Manchester, UK

We have developed, using text segmentation and keyword weighting methods, an information retrieval system for finding journal articles with information not directly alluded to in their title, abstract or supporting metadata. The system has been applied to a corpus of 21,000 relevant full text articles from the reproductive biology literature.

Using natural language processing (NLP) techniques, we are developing a search engine that aids the retrieval of documents holding fetal growth and development (FGD) information that is not alluded to in the title, abstract or supporting metadata of published mouse gene knockout studies. Mouse gene knockout studies produce countless observations of fetal and post-natal development, which are often documented in journal articles not directly addressing FGD, and are therefore likely to be missed by FGD researchers. Recovering such 'hidden' articles would give FGD researchers valuable leads for identifying novel mechanisms in fetoplacental development and function.

After studying FGD data and the searching strategies used by FGD researchers, text-mining techniques were applied to a sample set of mice knockout papers. This showed that the precision of keyword-based searching in scientific articles could be improved by using keyword-weightings for each section of discourse. An NLP method was created to extract the structured content from a HTML scientific article for representation and storage in a relational database. An established method for calculating keyword weightings was adapted to use the structural properties of articles so that keywords are weighted for each section of an article. A search form, allowing compound queries, was implemented as an interface to a fuzzy-Boolean querying system, which provides a more flexible method for scoring and ranking documents with weighted keywords than strict Boolean systems permit. These components were incorporated into the 'Mouse Knockout Article Search' ('MKAS') system. Results & Discussion: The system holds and analyses over 21,000 mouse gene knockout articles (collected using PubMed) from 1994 to 2004. We are now measuring the precision and recall of the document structure capturing process - early results show that the documents' structure and content are represented well in the database. The precision and recall of the entire search engine is then to be measured, the results of which may be used to change the weighting schemes for improved performance. The search mechanism currently ranks documents by their sections, so that the highest ranked document has the highest ranked section, the second highest ranking document has the second highest ranking section, and so on. The articles are mostly split into sections such as Introduction, Abstract, Methods and so on; consequently, the facility to navigate through such sections has been added to the 'Results' page.

The behaviour and performance of key areas of our system will be assessed. We aim to add further keyword-weighting strategies and NLP techniques to gain improvements in the system performance.