RDF CURATOR: A NOVEL WORKFLOW THAT GENERATES SEMANTIC GRAPH FROM LITERATURE FOR CURATION USING TEXT MINING

Yusuke KOMIYAMA[1], Osamu GOTOH[1]

[1]*Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Japan*

Background: There exist few databases that enable cross-reference among various research fields related to bioenergy. Cross-reference is highly desired among bioinformatics databases related to environment, energy, and agriculture for better mutual cooperation. By uniting Semantic Graph, we can economically construct a distributed database, regardless of the size of research laboratories and research endeavors.

Purpose: Our purpose is to design and develop a workflow based on RDF (Resource Description Framework) that generates Semantic Graph for a set of technical terms extracted from documents of various formats, such as PDF, HTML, and plain text. Our attempt is to generate Semantics Graph as a result of text mining including morphological analysis and syntax analysis.

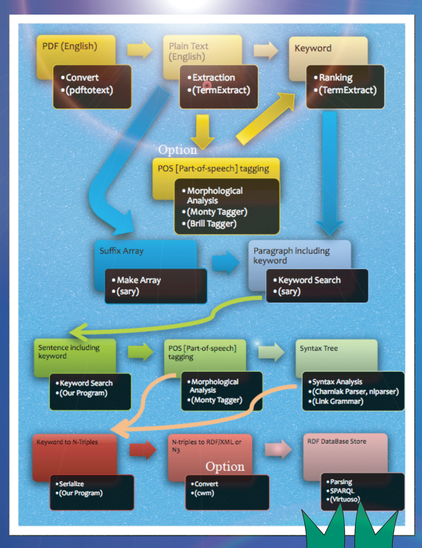

Figure 1 The work flow of *RDF Curator*. This system is composed of a graph of keywords that are extracted and registered in the DB automatically.

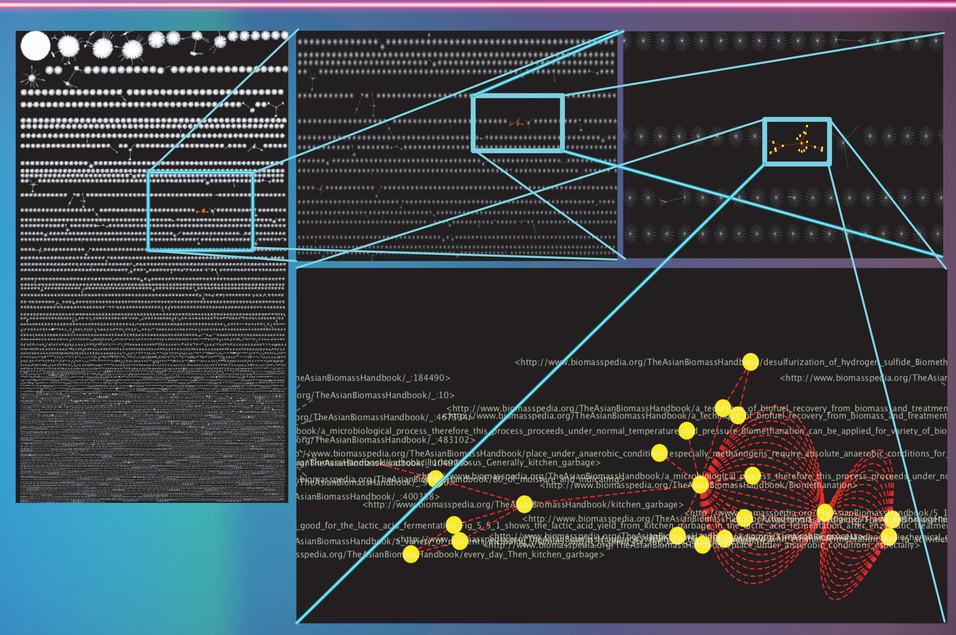

Figure 2 A sample result obtained by *RDF Curator*. A PDF document was serialized into RDF and then parsed to generate a relational graph database. Cytoscape was used as the visualization tool in this experiment.

Table 1 Numeric outline of data set for the test with *RDF Curator*.

Dataset	Pages of A4	Characters	Words	Num. of queries
The Asian Biomass Handbook	338	517,208	93,518	15,655

Table 2 A experimental result of data set for the test with *RDF curator*.

Dataset	Triples	Exe. Time (hour)	Num. of Nodes	Num of Edges
The Asian Biomass Handbook	135953	48	71766	64335

Result: We have developed a prototype of workflow program named *RDF Curator*. By using this system, various types of documents can be automatically converted into RDF. *RDF Curator* is composed of general tools and libraries so that no special environment is needed. Hence, *RDF Curator* can be used on many platforms, such as MacOSX, Linux, and Windows (Cygwin). We expect that our system can assist human curators in constructing Semantic Graph.

Conclusion: Although fast and high throughput, the accuracy of the present version of *RDF Curator* is lower than that of human curators. As a future task, we have to improve the accuracy of the workflow. In addition, we also plan to apply our system to analysis of network similarity.

Reference: Belleau F., et al., Bio2RDF: towards a mashup to build bioinformatics knowledge systems, *J Biomed Inform*, 41: (5)706-716, 2008.

Nature Proceedings : doi:10.1038/npre.2010.5072.1 : Posted 21 Oct 2010