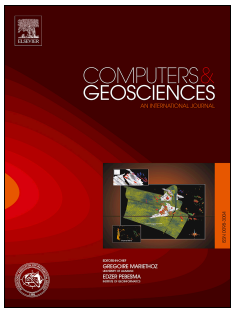
Accepted Manuscript



Information extraction and knowledge graph construction from geoscience literature

Chengbin Wang, Xiaogang Ma, Jianguo Chen, Jingwen Chen

PII: S0098-3004(17)30902-0

DOI: 10.1016/j.cageo.2017.12.007

Reference: CAGEO 4071

To appear in: *Computers and Geosciences*

Received Date: 29 August 2017

Revised Date: 10 December 2017

Accepted Date: 15 December 2017

Please cite this article as: Wang, C., Ma, X., Chen, J., Chen, J., Information extraction and knowledge graph construction from geoscience literature, *Computers and Geosciences* (2018), doi: 10.1016/j.cageo.2017.12.007.

Information Extraction and Knowledge Graph Construction from Geoscience Literature

Chengbin Wang [a, b], Xiaogang Ma[b*], Jianguo Chen [a*], Jingwen Chen [a]

[a] State Key Laboratory of Geological Processes and Mineral Resources & Faculty of Earth Resources, China University of Geosciences, Wuhan 430074, China

[b] Department of Computer Science, University of Idaho, Moscow ID 83844, USA

***** Corresponding Author E-mail: max@uidaho.edu (X. Ma); jgchen@cug.edu.cn (J. Chen)

Abstract: Geoscience literature published online is an important part of open data, and brings both challenges and opportunities for data analysis. Compared with studies of numerical geoscience data, there are limited works on information extraction and knowledge discovery from textual geoscience data. This paper presents a workflow and a few empirical case studies for that topic, with a focus on documents written in Chinese. First, we set up a hybrid corpus combining the generic and geology terms from geology dictionaries to train Chinese word segmentation rules of the Conditional Random Fields model. Second, we used the word segmentation rules to parse documents into individual words, and removed the stop-words from the segmentation results to get a corpus constituted of content-words. Third, we used a statistical method to analyze the semantic links between content-words, and we selected the chord and bigram graphs to visualize the content-words and their links as nodes and edges in a knowledge graph, respectively. The resulting graph presents a clear overview of key information in an unstructured document. This study proves the