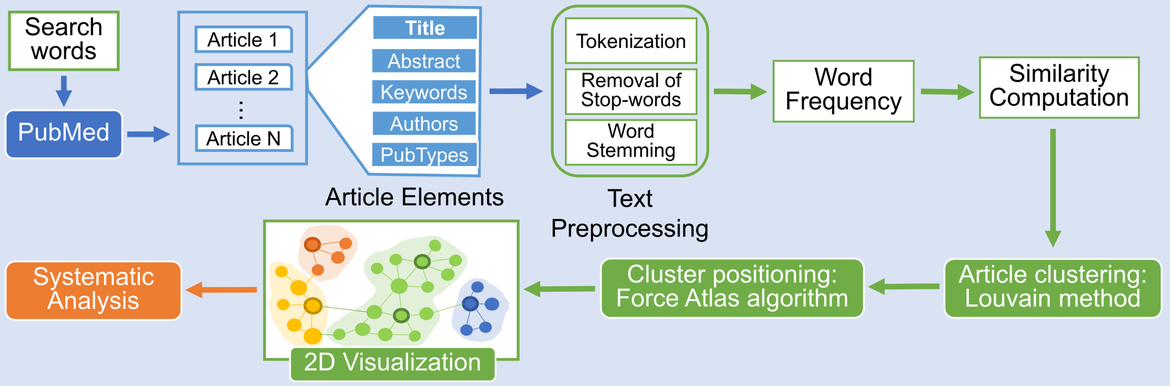
Data Mining from Collections of Scientific Papers: Illustrative Analysis of Groundwater and Disease

Yiding Zhang[1], Xiaonan Ji[2], Motomu Ibaraki[3], and Franklin W. Schwartz[3]
[1] Environmental Science Graduate Program, The Ohio State University, Columbus, OH 43210, USA
[2] Department of Computer Science and Engineering, The Ohio State University; [3] School of Earth Sciences, The Ohio State University

Introduction

The scientific research has been expanding exponentially for decades. Scientists are literally drowning in data and information. The goal of this paper is to demonstrate the power of data mining technique to evaluate large collections of papers and interpret the patterns of research strands. A systematic analysis of research on the emerging area of groundwater-related diseases was conducted as a demonstration.

Methods and Workflow



Clustering Results

- Search words: "groundwater", "disease"
- 426 research articles from 1971-2017
- 11 clusters were identified based on calculated article similarities
- The number of articles in each cluster ranged from 10 to 73
- Cluster topics were identified by keywords analysis

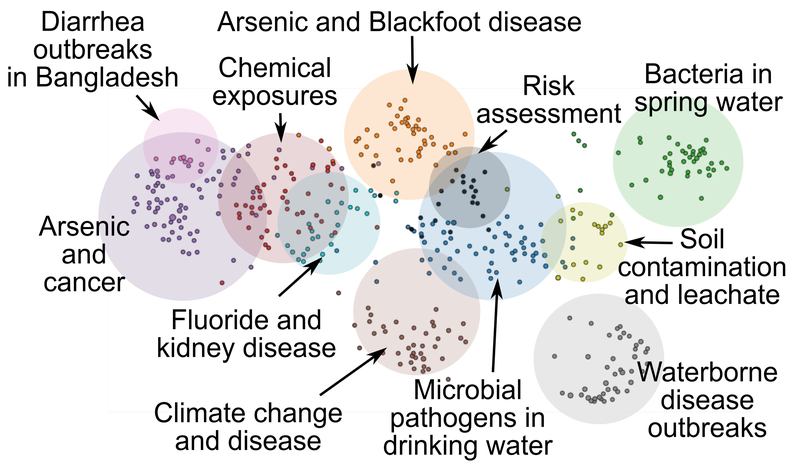


Fig. 1 Clusters and topics

Patterns of Diseases

- Two cluster groups were generated according to the contaminants-chemicals / heavy metals and pathogens
- The major diseases are cancer and diarrhea

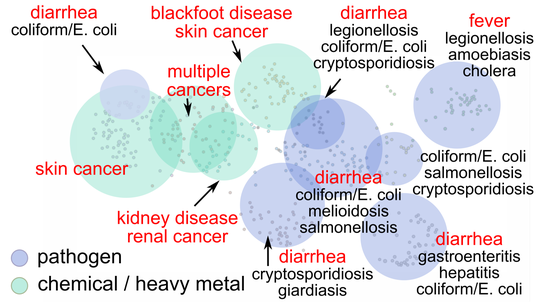


Fig. 2 Major diseases in the clusters

Groundwater vs Surface Water

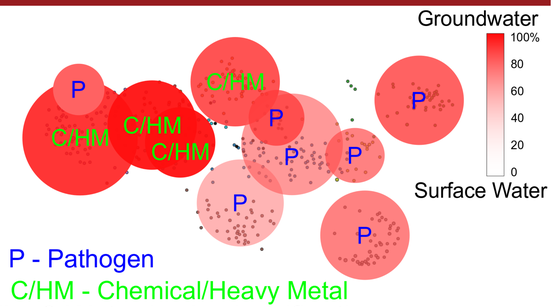


Fig. 3 Water body classification

- 88.3% of the retrieved articles are focused on "groundwater"
- Most chemical related articles study groundwater only; pathogen studies are closely related to surface water
- Only 7% papers mentioned groundwater chlorinated solvent contamination in health respects

Growth Patterns in Topics

- Articles focusing on chemicals grew slower before 2013, but they surpassed pathogen related articles in recent years
- Health and disease has emerged as a significant area topic relevant to groundwater since 2000

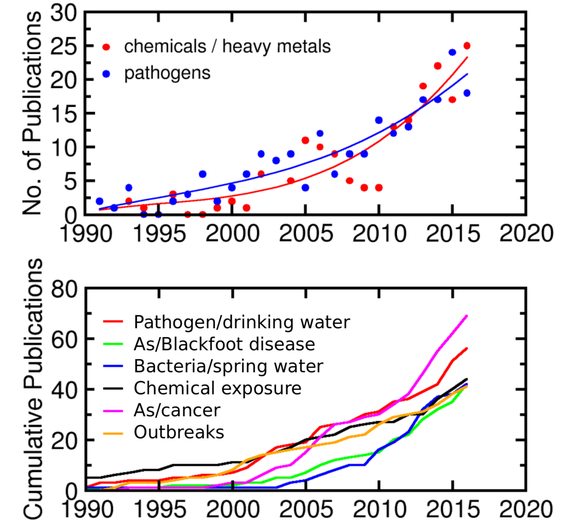


Fig. 4 Article growth patterns

Conclusions

- Research on water-related diseases in groundwater focuses mainly on chemicals and pathogens
- Cancer and diarrhea are two major diseases associated with chemical and pathogen studies, respectively
- Research on groundwater diseases is also related to surface water studies
- The 30 years efforts of chlorinated solvents studies in the U.S. appear not to be a significant driver for health-related research