MSCI 446 Data Mining Project Proposal

Project title: PaperScraper: Data mining approach to aid multi-document summarization for empirical literature review

Group Members:

Andrew Andrade {a2andrad, 20414374} Andy Toulis{aptoulis, 20466069} Baha Nurlybayev{bnurlyba, 20313597}

Abstract:

In short, we are looking to explore how data mining methodologies can be used to aid researchers in conducting a literature review. The objective of the project is to scrape the abstracts of existing publications and present users with domain specific preliminary analysis of concepts, ideas and tools. As a proof of concept, we will generate a literature review for the use of Artificial Intelligence and Predictive Analytics (AIPA) in the petroleum industry.

Previous work has been on multi-document summarization with a focus on autogeneration (which is difficult to do well especially for liturature view). Contrastingly, the goal is to utilize the insights gained from this approach for defining the scope of concepts to be summarized, not summarizing the concepts themselves. Manually searching, reading and reviewing relevant publications is highly time consuming. We are looking to aid researchers in finding points of interests in terms of relation, co-occurrence and time. We will be taking previously presented methods for document classification and clustering in a unique framework to create a useful tool for researchers review liturature.

Details

1. Source of data

For our proof of concept, we will be focusing on literature reviews in the petroleum industry with a specific focus on AIPA techniques used in Exploration and Production (E&P) of oil and gas. The dataset to be mined is publication abstracts and metadata scraped from OnePetro.org. The data will be scraped using the `mechanize` library in python, and cleaned using the

`beautifulsoup` library in python.

2. Attributes in the data set/ Schema/Number of rows:

OnePetro.org contains around 183,895 publications that we will be using as rows, and around 10 metadata columns per publication. The metadata will include the title, authors, publisher, source (specific conference, journal etc.), publication date, disciplines (based on Society of Petroleum Engineer's tags), keywords, number of downloads, file size and number of pages.

3. Use of data mining techniques:

Clustering:

We will be using a bag-of-words approach paired with a variation of k-means using a sparse matrix for clustering using either abstract text of publication or metadata of publication or both as features.

Classification:

We are evaluating our options in using Naive Bayes, Bernoulli models and SVM to use for classification. Our class variables will be defined as the E&P applications established based on domain expertise of a team member in the petroleum engineering field. The same features (publication abstract and metadata) will be used in classification as were defined in clustering.

Association Rule Mining:

We will likely be using Apriori for understanding associations of the AIPA techniques as the attributes. This will give insights to researchers on how techniques are paired togeather.

4. Prior Work and liturature search

Extensive research has been done and published in the field of multi-document summarization (Ladda Suanmali, Naomie Salim "Literature Reviews for Multi-Document Summarization") and some that focused auto-generation of summaries in the form of literature reviews (Kokil Jaidka,

Christopher Khoo, and Jin-Cheon Na, “Imitating Human Literature Review Writing: An Approach to Multi-document Summarization”). Our approach is different because instead of the replacing the researcher (summarizing the documents automatically), our focus is on aiding them (man-machine symbiosis) in reviewing a specific topic set. This is more revelant to researchrs as liturature review is currently done as a manual task.

5. Importance and Topicality

The idea is interesting because it can fundamentally change the way researchers do literature reviews. A data-driven methodology of reviewing a broad set of publications would provide analysis of the relation, trends and co-occurrence of concepts, ideas and tools. Connections which may not surface through manual searching will be captured automatically. In the petroleum engineering industry, researchers, engineers and management in the upstream side of oil and gas could benefit from the findings that we would be able to present in the E&P specific literature review that we intend on doing as a proof-of-concept.