Application of Text Mining for Enhanced Power Network Data Analytics – Offline Analysis of Textual Data

Introduction

Text mining is a process of condensing a text into a shorter version with only important information conserved. Being appreciably beneficial in many fields recently, text mining is expected to be applied in the future distribution networks. This poster is intended to explore the appropriate text mining methods for power system data analytics via practical applications. Focus is mainly on the first steps taken towards the knowledge discovery of large offline document collections.

Motivations

Due to increasing consumer demands and the integration of distributed energy resources, the present power distribution networks are of growing complexity and the volume of research studies to address different issues are considerably expanded. Enhanced data analytics for textual information could effectively allow the power system researchers to learn state of the art more efficiently. This as a result gives more comprehensive viewpoints and the overall operation and controls of power systems could be improved.

Methodologies

Research papers are available both online and offline. Online textual information requires web crawling and sometimes multilingual techniques as preparation. Offline analytics simply starts with a collection of scientific documents in pdf format. Figure 1 demonstrates the overall process flow.

The first step, text pre-processing, comprises converting the format from pdf to txt and eliminating textual interference such as headings and references. Next terms that have certain meanings are extracted and ranked according to their occurrence frequency. Term Frequency & Inverse Document Frequency (TF-IDF), the multiplication of the two independent measures, is the key to extract out terms that are representable for a certain topic. If TF is large, then the term is popularly used in the corpus; If IDF is large, then the term is only mentioned in few documents; if TF-IDF is large, the term is highly mentioned in few documents, in other words, these papers belong to the same category.

The second step creates a set of qualified terms as classifiers for categorisation. To enhance accuracy, human-beings are here involved to categorise chosen terms into various categories whose indexes are determined at the same time. After that, weightings are added manually for each term so that lexicons for different categories could be systematically established.

So far, terms become useful pointers and different text mining techniques can be applied. Here in the project query-based information extraction is undertaken in two forms: (1) Paper categorisation is to categorise documents in corpus into different topics and ranked accordingly. (2) Sentence summarisation basically is to group a few sentences exist in the papers to summarise the key findings in the corpus. Papers or sentences are ranked according to the terms they contained.

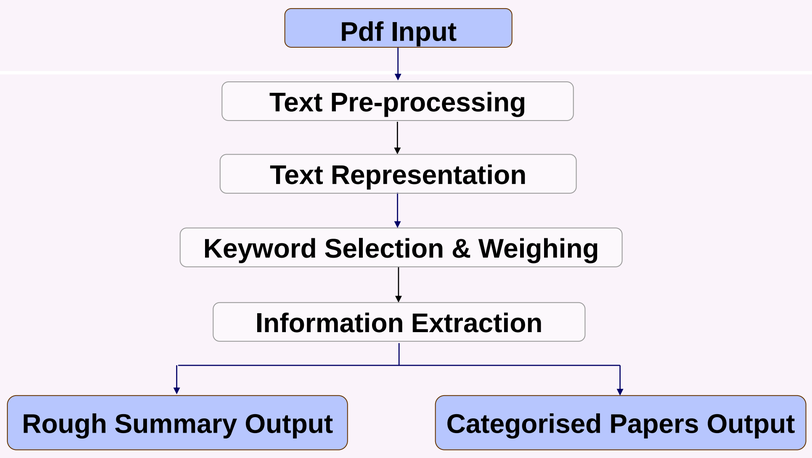

Figure 1 Flow Chart of the Text Mining Process

Results

In different steps, different software or techniques are used to transform the input into specific types of output. In the first step, text pre-processing has pdf document as input and main body of research paper in txt format as output. In the second step, as shown in Table 1, phrases/terms are ranked according to TF and their IDF and TF-IDF are calculated. The output results here as specialised terms did draw the range but they are not mature enough for enhanced knowledge discovery. Human supervision is therefore added as the third step.

Table 1 Results of Term Extraction

Ranking	Terms	TF	IDF	TF-IDF
1	power system	3149	0.2	780.3
2	reactive power	1923	0.5	902.3
3	active power	1359	-0.3	460.6
……	……	……	……	……
138	load curve	109	1.3	140.1
138	system load	109	1.5	133
139	reference voltage	108	1.2	159.4
140	load change	107	1.3	137.5
……	……	……	……	……
m	Term k	3	IDF_K	$TF\text{-}IDF_K$

Table 2 Results of Keyword Weighting

Demand Side Management	Weightings
demand response	5
demand profile	5
electricity demand	4
load demand	4
load profile	3
electricity consumption	3
actual demand	2
electrical load	2
residential load	1
local demand	1

Table 2 only shows ten of the terms selected for the category of demand side management. Based on the expert opinions, they are weighted and ranked according to their relevance to the topic. These terms form lexicons and their weights help enlarge the difference among papers or sentences.

Table 3 Results of Paper Categorisation

Paper	Topic 1	Topic 2	Topic 3
0001.txt	45	27	0
0002.txt	44	12	1
0003.txt	42	13	2
0004.txt	42	0	1
0005.txt	41	24	0
0006.txt	39	25	5

Table 4 Results of Sentence Summarisation

Sentence	Paper	Score
Sentence 1	0001.txt	15
Sentence 2	0002.txt	15
Sentence 3	0001.txt	13
Sentence 4	0003.txt	12
Sentence 5	0001.txt	12
Sentence 6	0002.txt	12

Table 3 and 4 are only segments of the final results. The real content of papers and sentences are represented by indexes. Differences among paper weightings are much larger then sentence's and therefore paper categorisation has higher accuracy of telling relevance then sentence summarisation.

Regarding the verification of the results, there are predefined evaluation criteria which assess papers or sentences with scores.

Conclusions & Future Work

In this project, a complete text mining process on offline scientific papers is developed within the framework of the traditional text mining applications. Given the currently available methods and tools, paper categorisation could be much more useful than sentence summarisation and the overall performance will be enhanced if more interference could be reduced. Further improvements should focus on the noise elimination and automated evaluation system should be developed for higher efficiency and accuracy.

Electrical Energy and Power Systems
School of Electrical and Electronic Engineering
The University of Manchester

For details contact: Yushi Chen
Email: yushi.chen@postgrad.manchester.ac.uk
Supervisor: Professor J.V.Milanovic
Email: jovica.milanovic@manchester.ac.uk