Towards application of text mining for enhanced power network data analytics - Part I

TOWARDS APPLICATION OF TEXT MINING FOR ENHANCED POWER NETWORK DATA ANALYTICS – PART I: RETRIEVAL AND RANKING OF TEXTUAL DATA FROM THE INTERNET

Piaoran Chen[1], Jelena Ponoćko[1*], Nikola Milošević[2], Goran Nenadić[2], Jovica V. Milanović[1]

[1]Electrical Energy and Power Systems Group, University of Manchester, Manchester, UK
[2] School of Computer Science, University of Manchester, Manchester, UK
**jelena.ponocko@manchester.ac.uk*

Keywords: Data analytics, text mining, web crawling, database

Abstract

This paper presents initial results of methodologies developed for extracting useful information from on-line textual data in the field of power networks. Since there are no tools developed specifically for power engineering, an attempt was made to take advantage of the existing tools for specialized web browsing and data extraction. Two methodologies are explored: one for extracting on-line documents (journal papers, technical reports, etc.) that are highly related to a specific topic, and the second one for extracting highly related sentences, providing a literature summary on the topic. The first results are promising, showing great potential for text-mining applications as a part of power networks data analytics.

1 Introduction

One of the most important distinguishing features of future power grid operation will be the increased use of communication and information technologies (ICT) in the generation, transmission and distribution part of power networks [1]. This will drastically increase the number of data sources and types of data used in the analysis of the network. Distribution networks will particularly change due to large-scale integration of low carbon technologies (LCT) and reliance on demand side management (DSM), i.e. active participation of the end-users in distribution network operation. With a higher involvement of customers, operation of the future distribution network should become more flexible, but also entangled with more uncertainties. In order to identify critical factors that could affect network operation, such as consumers' behavior or intermittent power generation from renewable sources, data mining methods (data analytics) have been increasingly developed in the power systems area.

One of the uprising information needs in distribution systems analysis is knowledge about customers' willingness to participate in DSM actions and conditions under which they would accept to change their daily habits in electricity usage. Next to the development of numerical data mining methods to retrieve this information, there is also an untapped potential among existing and publicly available textual data to find out these and similar pieces of information with the use of text mining methods. This data is usually given in tables or in the form of plain text in technical reports and research papers. Some authors have also pointed out the possibility of using Internet as a large database from which novel information can be extracted [2]. With tens of billions of pages [3] and an estimate of 966 exabytes (1 exabyte = 10^{18} bytes) of data exchanged annualy [4], Internet presents the world's largest database. Following this, more and more journal papers and articles are available on-line, which enables faster information exchange and update.

There is a vast research carried out in different parts of the world about power networks and plenty of projects are focusing on the DSM perspective. Many distribution companies have selected pilot sites to carry out trials, aiming at investigating operation possibilities and consumers' opinion about DSM. Although these trials are not large-scale, the corresponding findings are capable of providing basic knowledge in all aspects of DSM. Results of these projects are given in numerous reports and research papers, most of them being available on-line. This represents a valuable source of textual data accommodating different types of information about DSM.

Automated text mining methodologies have been widely used in medical and financial sectors to harness textual data on large-scale. Text mining (or natural language processing) is subfield of artificial intelligence [5], aiming to extract information from free text. Biomedicine, as one of the most mature areas of text mining technologies application, has demonstrated that text mining shows high effectiveness in revealing associated terms in order to identify causes of diseases and corresponding treatments [6-8].

This paper explores text mining methodologies in the area of power networks. We combine web crawling (traversing Internet to harvest documents of interest) and document ranking to identify documents and sentences of interest. We use the DSM as a case study

and demonstrate the initial results of web crawling of the on-line text data. To the best of our knowledge, this is a first attempt to collect such data in the power systems area.

2 Potential of Text Mining in Power Systems Area

Since there are no text-mining tools especially developed for crawling the power networks area, we explored the possibilities offered by existing tools for text mining or web data mining, with some modifications and refinement. As for web crawling, there are several applications developed such as Scrapy in Python [9], Nutch in JAVA [10] and Nokogiri in Ruby [11]. Since the task of designing a web-crawling tool for the electrical engineering area is not trivial, this paper aims to show what are the current possibilities based only on already available tools. One of the most efficient text-mining tools, IBM SPSS Modeller, has been widely used, providing a user-friendly interface and comprehensive data mining algorithms; however, these require significant customisation.

One of the greatest benefits of text mining is to minimise user's involvement needed for retrieval of useful information from Internet. It is pointed out in [12] that approximately only 6% of web pages have useful/educational content, which brings out the need for application of data mining methods that will eliminate advertisements and redundant websites. This could enable users to save a lot of time retrieving useful information from websites. Also, a series of research has been carried out to modify web mining results according to users' personal preferences [13] [14] [15], [16]. For example, it would be very time-consuming (or even impossible) for researchers to read all relevant papers about a specific topic as part of their literature review, since there might be thousands of related documents [17]. Therefore, text mining enables a filtering process which excludes documents that are not highly relevant to the field of interest, narrows down the number of documents and finally summarises the content of a smaller number of documents. Although the accuracy of this process is not guaranteed at the moment, initial results may show the usefulness and give direction for future research in this area.

This paper presents methodologies of web crawling and text mining to enhance power network data analytics. The first step towards discovery of useful information from textual data is to acquire adequate textual documents from Internet and then, in the following step, filter less relevant parts of the documents to obtain meaningful information. The methodology developed is illustrated in this paper using the topic of DSM, aiming to maximise the amount of information that can be obtained from Internet, as well as the quality of this information.

3 Data Acquisition via Web Crawling

In order to collect existing web pages on Internet into a database format convenient for further text mining, there is a need for web crawling. The general process for acquisition of on-line textual data is presented as a three-stage flowchart in Fig. 1. As shown in the figure, the quality of data acquired through web crawling is highly dependent on the selection of terms (key words or phrases) used in online searching. There are three different methods of generating search terms:

a) using a specifically defined topic as the keyword,
b) discussing with experts in the area of interest to generate a list of relevant terms,
c) obtaining relevant terms from a topic related paper by automatic term recognition (ATR)

We have explored all three approaches. For the ATR method, we used FlexiTerm [18], which recognises terms automatically in two steps: first it selects term candidates based on linguistic patterns, and then uses a measure called 'termhood' in order to identify relevant terms and discard common words. The output of FlexiTerm is a .csv format file listing all the terms appearing in given document(s) according to their term frequency (TF) and their 'termhood'. In this way, the most relevant mentioned terms can be identified and adopted for the following Internet search process.

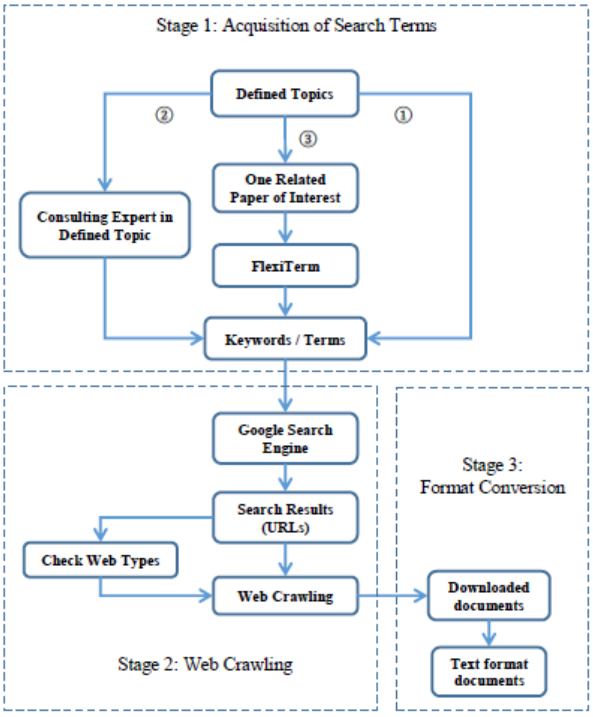

Figure 1 Database generation

After identifying relevant search terms using the different methods in stage 1, stage 2 includes crawling webpages to obtain documents related to the desired topic. For the first stage, the Eclipse programming environment is chosen for its compatibility with Python

programming language. Using Eclipse, the keyword list is first transmitted to a web scraping (data extraction) tool, which then uses Google Application Program Interface (API) to search for all websites related to given keywords, returning the top ranked webpages according to Google's comprehensive ranking. The result given by the scraping tool is a series of Uniform Resource Locators (URLs) and the number of URLs returned can be adjusted according to user's preference. The URLs related to the topic are saved in a text file for the web crawling process.

Crawling of web pages then includes downloading the content of the URLs. We also automatically downloaded documents that were directly addressed by URLs or series of documents addressed by links inside webpages. . As we are interested in research papers and project reports, we focused on pages containing PDF and PPT (PowerPoint) documents, since these on-line documents usually carry most information in the power networks area. Therefore, in stage 3, the resulting files in the PDF or PPT format were converted into text format as a necessary step before text mining. A script is written in Python based on 'PDFminer' [19] to automatically convert all documents into textual documents and numerate them successively for the convenience of tracking in the following text mining steps.

4 Text Mining Methodologies

In order to enhance information retrieval from webpages, two text mining methodologies are explored: one to rank documents resulting from the web crawling process, and one to extract important sentences, as shown in Fig. 2.

4.1 Ranking of Documents

Ranking of documents has an aim of finding documents that are highly related to a predefined topic in semi-automatic way, with as little of user's interaction as possible. The generation of a ranking table of terms (keywords) is one of the most significant steps in this process, as the ranking of documents is based on the quantity of relevant terms they contain. We considered two approaches for generation of the ranking table, both of them relying on the output of FlexiTerm. The first approach relies on experts' knowledge – manual ranking of terms/keywords resulting from the processing of all retrieved documents by FlexiTerm. The importance of keywords is identified according to specialists' experience in the corresponding field. Another approach is to rely on FlexiTerm's output after processing one highly relevant paper whose content matches with user's area of interest. This document is usually called a seed paper, and, in this case, a ranking list of keywords is obtained by manual selection of related terms from the list given by FlexiTerm. In both cases, the ranking list of terms should have around 50 terms.

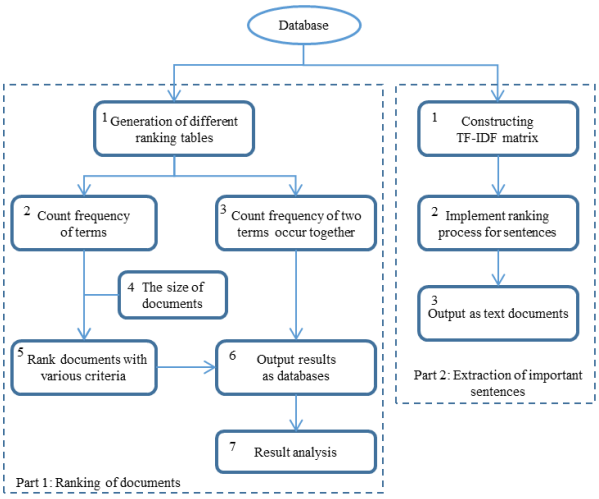

Figure 2 Document ranking end extraction of important sentences

There are three criteria combined to derive the final ranking of documents: the total number of occurrences of keywords (hits), weighting (importance) of the keywords, and the percentage of text containing keywords in each document. Considering these criteria together, the final output is a ranked list of all documents according to their total score.

4.2 Extraction of Important Sentences

Another methodology extracts important sentences from a series of documents, aiming to provide a summary of a given topic. A score is calculated for each sentence based on the number of keywords in the given sentence and the position of the sentence in the document. Depending on the number of sentences desired to be extracted, the output will be a text file containing the extracted sentences based on the ranking of their final score.

The first step towards sentences extraction is to presente term frequency in a form of a matrix called *Term Frequency - Inverse Document Frequency* (TF-IDF) matrix, constructed as illustrated in equations (1-5) [20]. Each document d_i ($i=1,..,m$) in a set of m documents is presented as a vector of n terms:

$$d_i = (f_1, f_2, \dots, f_n) \quad (1)$$

where f_j ($j=1,..,n$) is the frequency of the term t_j in the document. Therefore, the term-frequency matrix for the set of m documents is given as an $m \times n$ matrix:

$$D = (d_1, d_2, \dots, d_m) \quad (2)$$

In order to eliminate bias coming from long documents which as a consequence have higher frequency of the same term in respect to short documents, augmented normalised term frequency is used as follows [21]:

$$TF(t,d) = 0.5 + 0.5 * \frac{f_{t,d}}{max\{f_{t',d} : t' \in d\}} \quad (3)$$

where $f_{t,d}$ is the term frequency of term t in document d and $max\{f_{t',d} : t' \in d\}$ is the term maximum frequency of any term in the document. Since common words, such as "the" usually have high term frequency

but don't carry any information, the inverse document frequency (IDF) for term *t* with respect to database D is defined to represent the number of documents within database *D* that contain *t*, as illustrated:

$$IDF(t,D) = log_e \frac{m}{|\{d \in D : t \in d\}|} \quad (4)$$

where $|\{d \in D : t \in d\}|$ is introduced as *document frequency* (DF), representing the number of documents within the database that contain term t_j. Finally, TF-IDF matrix is obtained as follows:

$$TFIDF(t,d,D) = TF(t,d) * IDF(t,D) \quad (5)$$

The use of TF-IDF matrix will maximise the impact of a term when both conditions are met: high term frequency and low document frequency, i.e. small number of documents having high frequency of a specific term. Thus, due to the logarithmic characteristic of *idf-df* function, the common terms tend to have extremely low *idf* so that they are filtered out after the calculation of TF-IDF matrix.

In the following step, each sentence in the documents is given a score as follows [22]:

$$score = \frac{\sum_{t^i \epsilon s}^{m} tf-idf(t^i,d,D)}{\sum_{d^j}^{n} \sum_{t^i}^{M} tf(t^i,d) * tfidf(t^i,d,D)} \quad (6)$$

where *m* is the number of terms in sentence s that is also included in TF-IDF matrix, *M* is the total number of terms (columns) in TF-IDF matrix and *n* is the number of documents in the database. In order to include the impact of heading (or the first sentence) as indicator of the topic of a document, another score is added to the one defined in (6):

$$extra\ score = \frac{len(t)}{len(T)} * 0.1 \quad (7)$$

where *len(t)* is the number of term/terms shown in both the target sentence and heading/first sentence and *len(T)* is the number of terms shown in heading/first sentence and TF-IDF matrix.

The final step is to include the impact of sentence location, as adopted from [22]. The position P of a sentence is calculated as the line number of the sentence divided by the number of all lines in the text and presented in the range between 0 and 1. Based on Table 1 (adopted from [22]), another measure called distributed probability (DP) is obtained. The final score for each sentence is then calculated as follows:

$$final\ score = (score + extra\ score) * DP \quad (8)$$

Table 1 Distributed probability of important sentences (adopted from [22])

Position	0<P<0.1	0.1<P<0.2	0.2<P<0.3
Distributed probability (DP)	0.17	0.23	0.14
Position	0.3<P<0.4	0.4<P<0.5	0.5<P<0.6
Distributed probability(DP)	0.08	0.05	0.04
Position	0.6<P<0.7	0.7<P<0.9	0.9<P<1.0
Distributed probability (DP)	0.06	0.04	0.15

5 Case Study 1: Document Ranking

This case study illustrates an analysis done for the topic of DSM, as it is an important area of research in the field of power systems. Since DSM programs will rely on customers' willingness to participate, the aim of the study is to perceive customers' attitude towards DSM actions. For the illustration purpose, only two very broad search terms (phrases) are used, following the first method of generating search terms (Chapter 3), namely 'customer willingness to demand side management' and 'customer acceptance to demand side management'.

5.1 Construction of the Database

The web crawling process starts from executing the script developed to search for the two defined keywords (phrases) in the Google search engine, which gives as the output two text files, each containing 240 URLs (Fig. 3 illustrates a fraction of the results). Since only the URLs attached to PDF/PPT format files were considered in the following analysis, the number of documents downloaded via Scrapy, the web crawling tool, was only 195. If the search tool is set to save the first 240 searching results from Google database, it takes around 30 minutes to finish the searching and web crawling process for each defined keyword. Therefore, a selection of higher number of URLs is not time-efficient, which is why a smaller number of URLs was chosen, ensuring a satisfying quality of resulting documents at the same time.

```
10 http://www.a-eberle.de/en/product-groups/pq-mobile
11 https://www.mysciencework.com/publication/file/1500930/origin-of-
12 http://publications.lib.chalmers.se/records/fulltext/162855.pdf
13 http://www.fluke.com/Fluke/inen/Power-Quality-Tools/Logging-Power
14 http://en-us.fluke.com/products/power-quality-analyzers/fluke-174
15 http://www.jatit.org/volumes/research-papers/Vol8No2/3Vol8No2.pdf
16 http://arxiv.org/abs/1511.07281
17 http://ethesis.nitrkl.ac.in/6149/1/E-66.pdf
18 http://nrl.northumbria.ac.uk/344/
19 http://www.eweb.org/powerquality/report
20 http://vtb.engr.sc.edu/vtbwebsite/downloads/publications/01564207
```

Figure 3 Example of URLs resulting from Google search

In order to prepare documents for text mining as the next step, another script written in Python performed the transformation of PDF to text automatically and successively for all documents. In this case, there were 182 documents out of 195 which were converted into text files, since some documents failed at either web crawling or conversion process.

5.2 Generation of the ranking table of documents

In order to discover customers' opinion about DSM, a paper giving a literature review on this topic [23] was carefully chosen in order to use the second approach for the ranking table described in Chapter 4.1. After the automatic analysis of the paper using FlexiTerm, the corresponding term frequencies inside the given paper were ranked as presented in Table 2.

Table 2 FlexiTerm output for the given paper

Rank	Term representative	Score	Frequency
1	literature review of major trials	152.9673	142
2	domestic sector	111.3656	163
3	critical peak	110.3144	161
4	trial literature	97.9826	145
5	literature review of major trials overview	97.5299	72
6	literature review	97.3352	144
7	peak demand	87.3637	128
8	peak period	71.5583	105
9	peak reductions	70.8077	104
10	peak demand reductions	70.5858	66

The complete output of FlexiTerm for this paper had more than 500 terms. Considering the time needed to rank documents based on the number of terms chosen, it varied from about 30 min for 10 terms to about 5 hours for 500 terms. The 50 terms requiring about one hour to rank the documents were taken as the "optimal" for providing illustrative results for this pilot project.

(*Note:* It should be pointed out that web crawling and document ranking would take significantly less time than it was shown here, if the tools were used by trained computer science and text mining specialists. The study, instead, was performed by beginners in the area, i.e. electrical power engineers without any prior experience in text mining. One of the aims of this feasibility study was, after all, to investigate the feasibility and limitations of performing these studies by untrained professionals without any background in web crawling and text mining.)

5.3 Generation of the Output Document Ranking Database

The final output is an Excel document containing five worksheets: number of terms, existing keywords, number of sentences containing two keywords, records of related keywords and finally document ranking. The overall document ranking is the result of all aforementioned steps - all documents from the database constructed through web crawling are ranked according to the level of importance to the defined topic (Table 3).

Table 3 Overall ranking of the documents

Ranking of document	Document	Ranking of document	Document
1	Document11	11	Document78
2	Document1	12	Document2
3	Document31	13	Document64
4	Document46	14	Document27
5	Document37	15	Document26
6	Document43	16	Document59
7	Document21	17	Document45
8	Document3	18	Document10
9	Document81	19	Document169
10	Document55	20	Document70

Different colours are used to represent different levels of importance. While 'green' stands for strong relationship to given topic, 'blue' represents medium relationship and 'red' represents weak relationship to given topic.

In order to verify the accuracy of document ranking generated, the titles for 6 top ranked documents have been extracted in Table 4, to judge the actual relationship between the topic and each paper identified.

Table 4 Top ranked papers and reports

Paper Number	Paper Title	Number of Pages
paper 11	'Demand Side Response in the domestic sector - a literature review of major trials'	156
paper 1	Residential Demand Side Response for outage management and as an alternative to network reinforcement'	64
paper 31	'Smart Tariffs and Household Demand Response for Great Britain'	93
paper 46	'Assessment of Demand Response and Advanced Metering'	92
paper 37	'Developing the smarter grid: The role of domestic and small and medium enterprise customers'	50
paper 43	'The role of demand response in electric power market design'	57

Paper 11 [23] is the technical report chosen for extraction of keywords, as mentioned in Subchapter 5.2. It is focusing on the review of demand response (DR) trials in the UK and other countries as a means of reducing costs in power network and carbon dioxide emissions. It considers DR aimed at reduction of electricity usage during daily peak hours and exceptional critical peaks. Paper 1 [24] is another report discussing results of a trial which demonstrated that dynamic Time of Use (ToU) tariff can be a way to attract end-users to change their electricity consumption habits and provide benefits to the power network. Document 31 [25] is another report on smart meters' role in new electricity tariff design for consumers, including public policy, regulations and incentives. Paper 46 [26] presents a report on rollout of smart metering system in the United States and analysis of demand response potential across different regions. Report marked as paper 37 [27] covers customer flexibility in power consumption and generation, focusing on residential and small commercial users – it analyses flexibility of different home appliances and effects of heat pumps and electrical vehicles. Paper 43 [28] reports on the role of demand response in electrical market design, presenting different mechanisms and present issues.

As a conclusion, all highly ranked papers from the list are at the same time highly related to the topic of DSM and customers' involvement in demand response programs. They provide valuable information about users' flexibility from both technical and economic aspects. Further reading of these documents would certainly give deeper insight into the subject.

There are some terms that tend to show up together in one sentence in the documents. Table 5 presents pairs of highly related keywords that had the highest number of occurrence within the 182 documents. Following this, 'central terms' were extracted based on three criteria:

(i) The central term should show up with more than 10 terms recognised as related terms in worksheet 4 of the output database
(ii) The terms that are considered to have strong relation with the central term should occur more than 10 times
(iii) The resulting term pairs exist in more than one document (more than one paper given in third column of worksheet 4, as presented in Table 6)

Table 5 Worksheet 4 from the output database

Related keyword	Frequency	List of related papers
domestic sector AND literature review	152	paper2,paper11,paper15, paper26
literature review of major trials AND domestic sector	150	paper2,paper11,paper15
literature review of major trials AND literature review	150	paper2,paper11,paper15
critical peak AND critical peak pricing	40	paper1,paper3,paper11, paper21, paper26,paper27,paper31, paper37, paper46,paper48,paper61, paper64, paper81
peak demand AND demand reduction	38	paper3,paper11,paper21, paper25, paper31,paper43,paper45, paper55, paper65,paper71,paper81, paper160, paper162
critical peak AND peak period	25	paper3,paper11,paper21, paper27, paper31,paper45,paper55, paper81
demand response AND dynamic pricing	23	paper1,paper3,paper27, paper43, paper46,paper51,paper59, paper81, paper154
critical peak AND peak price	22	paper2,paper3,paper11, paper21, paper31,paper61,paper78, paper81
peak demand AND peak demand reductions	18	paper3,paper11,paper25, paper31, paper43,paper55,paper81
peak demand reductions AND demand reduction	18	18paper3,paper11,paper25, paper31, paper43,paper55,paper81

Therefore, in this case study, there are four central terms identified ('critical peak', 'peak demand', 'demand reduction' and 'electricity demand') and relationships with their corresponding strongly related terms are shown as an overview in Figure 4. For an unexperienced user, this could be a brief introduction to DSM illustrating the inner-relations between terms that are of critical relevance to DSM area.

With the ranking of documents, as well as generation of term relationships, documents closely related to a given topic can be identified from a large database. In case one wishes to obtain as much information as possible without reading all documents in the database, term relationships can provide a proper overview of the defined topic, identifying keywords, i.e. topics that should be analysed.

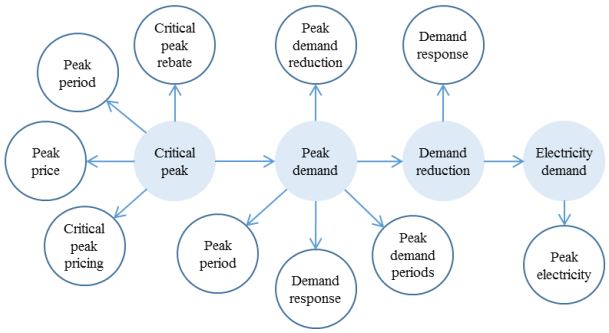

Figure 4 Term relationships for the topic 'demand side management'

6 Case Study: Sentence Extraction

The second case study presents sentence extraction methodology applied to the topic of 'power quality'. The database in this case consists of 783 documents obtained via web crawl. With the methodology for allocating score to each sentence, considering the value TF-IDF, as well as the location of the sentence, a text file containing 30 sentences with higher scores within the database is generated. For illustration purposes, there are four sentences with highest scores presented in Table 6 to show the quality of extracted sentences.

Table 6 Extracted sentences

paper 322	Conductive disturbances in the supplying voltage may also affects the root mean square voltage and shape of the voltage curve on the mains which reflect in degradation of the power quality.
paper 342	Introduction: The power quality problem is now of a great concern to electric utilities of power industry and they are trying hard to supply their customers with a good quality of power especially in the open market.
paper 71	Conclusion: The simulation approach provides the researcher the flexibility to create power system models to simulate power quality disturbance by connecting various functional building blocks in the simulation environment.
paper 342	Today, a new factor, harmonics, has been added to the power quality scenario because utility customers, including residential ones, are using electronic devices that require non-sinusoidal currents, currents rich in harmonics.

In these four sentences, it is stated that power electric utilities are trying their best to maintain high standard power quality, especially in open market, and thus they have to take into consideration the impact of harmonics and conductive disturbances in voltage supply. At the moment, the information provided by these highly ranked sentences is not logically and semantically arranged, since it is still at the stage of sentence extraction. However, this tool is capable of providing an overall idea of the topic after reading all sentences extracted. Another option for the user is to easily track from an informative sentence back to its original source (paper) and obtain more detailed information without reading other documents in the database.

7 Conclusion

This paper presents an initial investigation of the possibilities and challenges of applying existing web crawling techniques and text mining tools to enhance power network data analytics. Web crawling is used as the first step to generate a database from textual data located in numerous websites. Following this, two methodologies are developed: document ranking and

sentence extraction, former one to extract highly topic-related documents, and the latter one to generate a summary on a topic.

The most significant contribution of the developed methodologies is their ability of acquiring a series of documents or sentences that relate to defined topic. Although many widely used search engines have the ability to return millions of results according to users' queries, the noise level is not acceptable for power network research and, in addition to that, manual extraction of useful information from these results would be highly time consuming. Therefore, at this stage, the methodologies for text mining described in this paper present a highly useful novelty in advanced data analytics in power network area.

Future work will focus on developing existing methodologies towards more comprehensive results giving meaningful summaries of numerous on-line text documents and less need for user's involvement in the overall web crawling and text mining process.

8 Acknowledgements

This research is partly supported by the EU Horizon 2020 project "Nobel Grid", contract number 646184.

9 References

[1] A. Srivastava. *Cyber Infrastructure for the Smart Grid, [Online]*. Available: https://sgdril.eecs.wsu.edu/files/files/Lesson%201_2_Motivation%20for%20the%20Smart%20Grid.pdf

[2] M. Craven, A. McCallum, D. PiPasquo, T. Mitchell, and D. Freitag, "Learning to extract symbolic knowledge from the World Wide Web," DTIC Document1998.

[3] *How Big is the Internet*. Available: http://techland.time.com/2011/06/01/how-big-is-the-internet-spoiler-not-as-big-as-itll-be-in-2015/

[4] *World Wide Web Size [Online]*. Available: http://www.worldwidewebsize.com/

[5] V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," *Journal of emerging technologies in web intelligence*, vol. 1, pp. 60-76, 2009.

[6] D. R. Swanson and N. R. Smalheiser, "An interactive system for finding complementary literatures: a stimulus to scientific discovery," *Artificial intelligence*, vol. 91, pp. 183-203, 1997.

[7] D. Kumar and D. Bhardwaj, "Rise of data mining: current and future application areas," *IJCSI International Journal of Computer Science Issues*, vol. 8, 2011.

[8] M. J. Embrechts, B. Szymanski, and K. Sternickel, "Introduction to scientific data mining: Direct kernel methods and applications," *Computationally Intelligent Hybrid Systems: The Fusion of Soft and Hard Computing*, pp. 317-362, 2004.

[9] *Scrapy [Online]*. Available: http://scrapy.org/

[10] *Nutch [Online]*. Available: http://nutch.apache.org/

[11] *Nokogiri [Online]*. Available: http://www.nokogiri.org/

[12] S. Lawrence and C. L. Giles, "Accessibility of information on the web," *Nature*, vol. 400, pp. 107-107, 1999.

[13] J. Cho and H. Garcia-Molina, "The evolution of the web and implications for an incremental crawler," 1999.

[14] C. Chen, "Structuring and visualising the WWW by generalised similarity analysis," in *Proceedings of the eighth ACM conference on Hypertext*, 1997, pp. 177-186.

[15] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, pp. 604-632, 1999.

[16] C. Chen, J. Newman, R. Newman, and R. Rada, "How did university departments interweave the Web: A study of connectivity and underlying factors," *Interacting with computers*, vol. 10, pp. 353-373, 1998.

[17] G. B. Treasury, *The green book Appraisal and evaluation in central government: Treasury guidance*: Norwich (United Kingdom): TSO, 2003.

[18] I. Spasić, M. Greenwood, A. Preece, N. Francis, and G. Elwyn, "FlexiTerm: a flexible term recognition method," *Journal of Biomedical Semantics*, vol. 4, pp. 1-15, 2013.

[19] *PDFminer [Online]*. Available: https://pypi.python.org/pypi/pdfminer/

[20] C. D. Manning, P. Raghavan, and H. Schütze, "Scoring, term weighting and the vector space model," *Introduction to Information Retrieval*, vol. 100, pp. 2-4, 2008.

[21] A. K. Singhal, "Term weighting revisited," Cornell University, 1997.

[22] Y. Seki, "Sentence Extraction by tf/idf and position weighting from Newspaper Articles," 2002.

[23] F. Economics and S. First, "Demand side response in the domestic sector–a literature review of major trials," *Final Report, London, August*, 2012.

[24] "Residential Demand Dide Response for Outage Management and as an Alternative to Network Reinforcement," UK Power Networks Holdings Limited,2014.

[25] G. Owen and J. Ward, "Smart tariffs and household demand response for Great Britain," *Sustainability First, London*, p. 2010, 2010.

[26] F. E. R. Commission, "Assessment of demand response and advanced metering," 2008.

[27] J. Bird, "Developing the smarter grid: the role of domestic and small and medium enterprise customers," Customer-Led Network Revolution,2015.

[28] S. Braithwait and K. Eakin, "The role of demand response in electric power market design," *Edison Electric Institute,* 2002.