Extracting Distribution Network Fault Semantic Labels from Free Text Incident Tickets

Bruce Stephen, *Senior Member IEEE,* Xu Jiang, *Student Member IEEE*, and S.D.J. McArthur, *Fellow IEEE*

***Abstract*—Increased monitoring of distribution networks and power system assets present utilities with new opportunities to predict and forestall system failures. Although automated pattern recognition methodologies have given other industries significant advantage, power system operators face additional challenges before these can be realized. The effort of apportioning ground truth to fault data creates a knowledge bottleneck that can make utilizing automatic classification techniques impossible. Surrogate approaches using operational process outputs such as maintenance tickets as labels can be challenging owing to the causal ambiguity of these written records. To approach a solution, this paper demonstrates utilizing natural language processing techniques to disambiguate the free text in maintenance tickets for onward use in supervised learning of fault prediction and classification techniques. A demonstration of this approach on an established power quality fault data set is provided for illustration.**

***Index Terms*—Fault Diagnosis, Document Topic Models, Distribution Networks**

I. Distribution Fault Prediction and Classification

DISTRIBUTION Networks' observability has increased in recent years with the advent of low cost, high resolution monitoring devices. This has allowed network operators to capture the characteristics of fault signatures that would have previously gone unnoticed until they resulted in failure or outage [1]. The next logical step of this evolution of distribution network monitoring would be the automated identification of such faults, following other industries pursuit of leveraging data to enhance operation and understanding. The barrier to this is that in order to classify such faults, a set of labelled faults are required as exemplars in the first instance. Explicit labelling is time consuming and requires expertise to identify and articulate fault taxonomies, which may not reside within the business.

Ticket based maintenance records and directives exist in a number of service and infrastructure industries; for example, in [4], topic based models were used to understand the underlying problems from unstructured ticket text. In distribution network operation, often attached to faults are incident or maintenance tickets submitted for validation or work scheduling purposes. These too are typically free text, with a description provided by the individual who filed them and as such will not contain standardized terms or descriptions; instead, it will contain the perspective of the filing individual making it susceptible to ambiguity and hence unusable for supervised machine learning of fault diagnoses. One well-curated example of such a set of incidents though, is the EPRI/DoE National Database of Power System Faults [2], 13 examples of which are described in Table I.

TABLE I
DoE/EPRI Power Quality Fault Library: First 13 Records

Event Id	Phase	Cause	Weather	Details (free text)
0001	2	Tree	Clear Weather	Fault caused line recloser lockout. Tree Outside Right of Way (Fall/Lean On Primary)
0004	2	Tree	Clear Weather	Fault caused line recloser lockout. Tree Outside Right of Way (Fall/Lean On Primary)
0005	2	Tree	Clear Weather	Fault caused line recloser lockout. Tree Outside Right of Way (Fall/Lean On Primary)
3042	4	Equipment	Unknown	Equipment, Device UG, Damaged.
0021	1	Equipment	Clear Weather	Overhead Insulator Failure. BROKEN INSULATOR
0022	1	Equipment	Clear Weather	Overhead Insulator Failure. BROKEN INSULATOR
0062	4	Undetermined	Raining	storm
0064	4	Undetermined	Raining	storm
0067	4	Tree	Thunderstorm	Tree/Limb Growth
0065	4	Tree	Thunderstorm	Tree/Limb Growth
0068	2	Tree	Clear Weather	VINES ON TRANSFORMER
2760	1	Unknown	Unknown	Short duration variation. No outage information found.
3048	3	Equipment	Unknown	Equipment, Capacitor Station, Damaged.

This data set is unique in that it provides both the maintenance report ('Details') as free text as well as a ground truth classification ('Weather', 'Cause'); operationally, providing these classifications would be an unfeasible effort, so a means of automatically inferring these from the routinely available maintenance notes would be valuable, and this data set provides the means of validating such a method.

II. Document Topic Models

Without a semantic model, the labelling of fault occurrences using selected keywords from maintenance tickets would be

Dr. B. Stephen is a Research Fellow in the Advanced Electrical Systems Research Group, Institute of Energy and Environment, University of Strathclyde, Glasgow, G1 1RD (phone: +44 (0)141 444 7260, e-mail: bruce.stephen@strath.ac.uk)

prone to spelling, grammatical, style and terminology aberrations which could only be overcome by enforcing strict maintenance reporting guidelines which provides an additional burden on the field operative. A representation popular in the Natural Language Processing and Information Retrieval communities for many years, the 'bag of words' is highly suited to incident tickets and operative fault reports [3]: this entails 'stopping' the document (ticket) by removing common words, stemming all verbs and adverbs (which turns them into a corresponding noun) and leaves the document as a vector of word occurrence counts. This approach yielded a number of widely used document similarity metrics based on distances between these vectors that reflected commonality of terms. Subsequent probabilistic formulations of this approach could be used to imply polysemy and synonymy among terms making them ideal for identifying documents with the same sentiment but different term usage [5] – a characterizing problem of maintenance reports. Latent Dirchlet Allocation (LDA) [6] was one such model that represented a vocabulary of N words over a corpus of documents D:

$$p(D|\alpha,\beta)=\prod_{d=1}^{M}\int p(\theta_d|\alpha)\prod_{n=1}^{N_d}\sum_{z_{dn}}p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)d\theta_d \quad (1)$$

LDA builds a probability distribution of k topics z within a document, with each topic itself having a probability distribution of words w. These k topics, essentially 'cluster' variables constraining the choice of word distribution in a document, are purely hypothetical and are not encoded in documents explicitly. The use of a Dirichlet distribution with parameter α, selects the proportion or composition θ of topics in a given document:

$$P(\theta|\alpha)=\frac{\Gamma\left(\sum_{i=1}^{M}\alpha_i\right)}{\prod_{l=1}^{M}\Gamma(\alpha_l)}\prod_{j=1}^{M}\theta_j^{\alpha_j} \quad (2)$$

While β similarly parameterizes a conditional Dirichlet distribution of words over each topic. The dominant topic implied by a ticket may be used as an alternative to explicitly labelling fault cause – each topic ranks the words most likely to have generated them, thus providing a human readable interpretation.

III. LDA Topic Model of Exemplar Distribution Fault Incident Tickets

Figure 1 shows how the topic distribution would be generated and then associated with fault records – the resulting fault would be automatically labelled using the most probable terms for a given topic. Specific categories will be evident from the most likely word stems. As an example from another field, [5] demonstrated that a 'budgets' topic was found to generate words such as 'provide', 'facilities', 'foundation', 'fund' and an 'arts' topic generated 'performing', 'act', 'music', 'leading' and 'supporter' amongst others. Table II demonstrates the 10 most probable terms for each of 5 topics learned. The number of topics were chosen for brevity of illustration although formal selection procedures can be used to find the implied number of topics for an LDA model of a data set.

TABLE II
Latent Dirichlet Allocation 5-Topic Model

Topic	10 Most Probable Word Stems
1	caus, undetermin, breaker, primary, lightn, tree, storm, trip, fault, investig
2	lightn, transform, caus, outag, line, limb, tree, territory, substat, sag
3	ug, damage, equip, cabl, hit, connector, pole, vehicle, dig, car
4	tree, motor, pole, vehicle, oh, primary, fuse, right, way, fall
5	caus, trip, line, breaker, event, substat, reclos, unknown, time, transmiss'''

Table II shows that fault case specific categories are evident from the most likely word stems: topic #1 broadly corresponds to weather related events, topic #2 to vegetation encroachment (including those invoked by weather), topic #3 to conductor damage including cable and 3rd party related incidents (i.e. 'dig' and 'car'), topic #4 to conductor related faults such as tree impacts; topic #5 to equipment failure or transient faults. Further post-processing by a domain expert would identify 'ug' as 'underground' and 'oh' as 'overhead' which would overlay additional context. The nature of these faults cuts across multiple categories as the cause may be multifactorial, but accommodating this is a key feature of the LDA model [6].

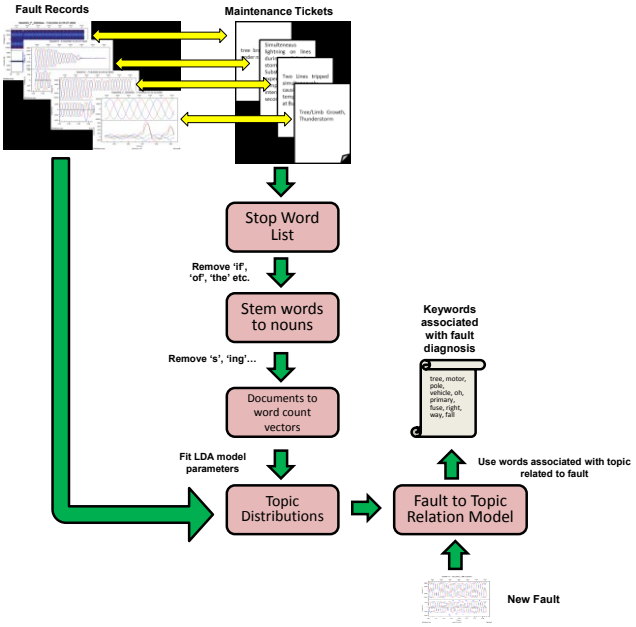

Fig. 1. Process for automatically labelling faults with maintenance records.

IV. Predictive Power of Topic Models

The key barrier to applying supervised machine learning techniques for fault diagnosis in power systems applications is the effort required to produce a sets of labelled exemplars for models to learn from. This section demonstrates how a topic model can be learned from a set of labelled maintenance tickets, the topics have a human interpretable form provided by the most probable words for the topic and that the topics can be used to predict the cause associated with the fault maintenance ticket. If there is sufficient predictive power in the maintenance ticket topics, then maintenance tickets have the potential to be used to label faults unambiguously and provide an automated form of

generating diagnostics to enhance fault situational awareness. Since maintenance tickets are produced under normal operational procedure, this removes the bottleneck associated with translating domain knowledge into machine learned profiles without the need for manual labelling.

The DoE fault data set already has a set of labels corresponding to the circumstances surrounding the recorded fault – if the maintenance ticket semantics can be demonstrated to relate to these, then in practice the ticket could be used to categorize a fault with any polysemous and synonymous terms accommodated by a topic model. To test this, a small selection of state of the art classifiers were trained to demonstrate if there was a relation between topics inferred from maintenance tickets and expert labeling [8]. Each maintenance ticket is converted to the bag of words representation and then the resulting word vector is run through the pre-trained LDA model described in Section III to get a topic vector associated with each fault record and its label. For predicting the expert label from just a topic vector, Table III shows the accuracy (the ratio of true positives plus true negatives to all classifications made) of 10 classification models, all of which work on different discriminatory principles and decision surface shapes.

TABLE III
CLASSIFICATION PERFORMANCE OF 5-TOPIC LDA MODEL

CLASSIFIER	MAINTENANCE TICKET LABEL PREDICTION ACCURACY
Ada Boosted Tree	54.7%
Decision Tree	76.2%
Gaussian Process	61.9%
Linear Support Vector Machine	59.5%
Naive Bayes	45.2%
Nearest Neighbor	**78.6%**
Feedforward Neural Network	69.0%
Quadratic Discriminant Analysis	45.2%
Radial Basis Function SVM	66.6%
Random Forest	73.8%

Using a 25% held out set from a selection of 168 labelled examples, Table III shows that given an appropriate classifier choice, the topic composition vector provided by the LDA model can be related, and is therefore implicit of the sentiment conveyed in the maintenance report since it corroborates with the label provided by the domain expert in the DoE data set. Since this relation has been shown to exist, the label can be replaced with a human readable description generated by taking the most likely topic words associated with a topic, as demonstrated in Table II.

V. ANTICIPATED PRACTICAL APPLICATION

This paper has proposed a means of automatically labelling power system faults by modelling the semantic content in maintenance tickets. In the operational environment described in Figure 1, this would allow digital fault records e.g. [2,9] that were associated with a particular network incident to be automatically labelled, using the semantic content of an accompanying maintenance report. An LDA model is used to produce a topic vector probability, *p(z)* from the report, which can be used along with its word/topic probability distribution as exemplified in Table II to generate a human readable label by choosing a subset of words that produce maximum values of

$$p(w_{nd}) = p(z_d)p(w_n|z_d) \quad (3)$$

Assigning descriptive text in this manner would deal with the bottleneck associated with producing training data for supervised learning of fault classifiers – with a readable description associated with a fault records, there would be no need for engineers to manually label exemplars.

Performance of around 75% for predicting fault cause from inferred document semantic content suggests that LDA models need larger corpora to learn from: LDA as originally formulated does not lend itself to learning word distributions from short documents i.e. maintenance tickets. Rather than imposing verbosity limits [10] and language guidelines on the filing of maintenance reports, an LDA model instead may be pre-trained on semantically related documents such as maintenance manuals or abstracts. Without a strategy for automation, power systems data acquisition systems [1, 2, 9] will continue to face the barriers associated with ground truthing fault diagnostic systems, will scale poorly to adoption as business as usual and will be incapable of unlocking the potential situational awareness that could be obtained through investment in infrastructure monitoring.

VI. REFERENCES

[1] Wischkaemper, J.A., Benner, C.L., Russell, B.D. and Manivannan, K., 2015. Application of Waveform Analytics for Improved Situational Awareness of Electric Distribution Feeders. IEEE Transactions on Smart Grid, 6(4), pp.2041-2049.

[2] DOE/EPRI National Database Repository of Power System Events, [Online].Available: http://pqmon.epri.com/disturbance_library/see_all.asp

[3] Passonneau, R.J., Rudin, C., Radeva, A. and Liu, Z. "Reducing Noise in Labels and Features for a Real World Dataset: Application of NLP Corpus Annotation Methods." In CICLing, pp. 86-97. 2009.

[4] Becker, H., and Arias, M. "Real-time ranking with concept drift using expert advice." In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 86-94. ACM, 2007.

[5] Hofmann, T., "Probabilistic latent semantic analysis," Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence, pp. 289-296, 1999.

[6] Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), pp.993-1022.

[7] Zhou, W., Tang, L., Zeng, C., Li, T., Shwartz, L. and Grabarnik, G.Y., 2016. Resolution recommendation for event tickets in service management. IEEE Transactions on Network and Service Management, 13(4), pp.954-967.

[8] Li, K., Xie, J., Sun, X., Ma, Y. and Bai, H., 2011. Multi-class text categorization based on LDA and SVM. Procedia Engineering, 15, pp.1963-1967.

[9] Strachan, S.M., McArthur, S.D., Stephen, B., McDonald, J.R. and Campbell, A., 2007. Providing decision support for the condition-based maintenance of circuit breakers through data mining of trip coil current signatures. IEEE Transactions on Power Delivery, 22(1), pp.178-186.

[10] Kenter, T. and De Rijke, M., 2015, October. Short text similarity with word embeddings. In Proceedings of the 24th ACM international on conference on information and knowledge management (pp. 1411-1420).