Tool for Debugging Transformations of Heterogenous Outage Data

Martin Hyndrich[1] and Michal Krátký[1] and Radomír Goňo[2]

[1] Department of Computer Science
[2] Department of Electrical Power Engineering
FEECS, VŠB – Technical University of Ostrava,
17. Listopadu 15/2172, 708 33 Ostrava-Poruba, Czech Republic
{michal.kratky, radomir.gono}@vsb.cz

Abstract. Knowledge of the component reliability parameters in power networks is necessary for the reliability computation of wholesale-consumer connection. Component reliability parameters are possible to retrieve only with accurate databases of distribution companies. Such a database includes records of failures and outages in power networks. The main issue for an analysis of these databases is the heterogeneity feature: we need a framework for a unification of various distributors's databases. In this case, data of various distributors is automatically transformed in the unified database. In previous papers, we introduced such a framework. Since this framework utilizes automatic transformations of outage records, we need to provide a tool for debugging of the transformations. In this paper, we describe this tool and we show that it enables us to minimize errors in transformations.

Key words: outage database, reliability computation, framework for data unification, debugging tool for data transformation

1 Introduction

It is necessary to observe failures and outages in the transmission and distribution of electrical energy[1] for retrieving the component reliability [2]. Furthermore, electrical energy unsupplied to consumers is possible to compute. Probability computation of unsupplied energy is only possible on the basis of reliability computation results.

Databases of power faults have been collected in the Czech republic for several years [6, 7]. A statistical significance of an outage database depends on the number of records in the database. A larger database would describe the

⋆ This work was supported by the Czech Science Foundation (No. GA102/09/1842) and by the Ministry of Education, Youth and Sports of the Czech Republic (SGS, No. SP2011/193).

[1] We have used the term outage database instead of the perfect phrase database of failures and outages in the transmission and distribution of electrical energy in this paper.

real condition of network equipment more accurately. Therefore, it is necessary to merge databases of various distributors. The main problem of the merging is the heterogeneity feature: databases of various distributors differ from one another. It is impossible to retrieve reliability parameters from this data in a direct way. For this reason, it is necessary to develop a framework for retrieving parameters from various databases. In [9, 8], we introduced this framework for retrieving reliability parameters in distribution networks. Since this framework utilizes automatic transformations of outage records, we need to provide a tool for debugging of the transformations. In this paper, we describe this tool.

This paper is organized as follows. In Section 2, we put forward a background of a reliability computation and outage database. In Section 3, framework for the outage databases unification is briefly described. In Section 4, we introduce the debugging tool and we show that it enables us to minimize errors in transformations. In the last section, the paper content is resumed and the possibility of a future work is outlined.

2 Outage Database

2.1 Outage Monitoring History

Research on power system reliability first appeared in the 1940s in the U.S.A., and later in the USSR and Great Britain. Since the 1950s, reliability research has carried on in all developed countries.

Component failure rates tend to vary with a component work life [10, 1]. Many parameters in the field of reliability vary for a specific component and the condition in which a component works. These random variables are represented by probability distribution functions [4].

Since, in this paper, we have applied SOM to the analysis of outage data in the Czech and Slovak republics, we should introduce a background of outage monitoring in these countries. In the former Czechoslovakia, discussions on power system reliability dated back to the 1960s. The turning point for reliability monitoring was in 1974: Regulations 2/74 for electric power systems CEZ[2] and SEP[3] were released [11]. These regulations unified failures, outages and damaged equipment monitoring options for all distribution companies in Czechoslovakia. Since 1975, exclusive outage databases have been on the rise.

This database is a very valuable baseline for reliability computation. Unfortunately, database building has ceased since 1990 because of political and social changes. Separate distribution companies have introduced their own systems for reliability monitoring since the 1990s. A complete database has not been built henceforth.

The expert group, CIRED Czech[4], has introduced a discussion on reliability issues. The first calls for integration of particular outage databases were already

[2] Czech energy company
[3] Slovak energy company
[4] `http://www.ckcired.cz/`

claimed at the first meeting of this group in 1997. In 1999, distributors opted for unified monitoring of global reliability indices and the reliability of selected pieces of equipment. Data for the reliability computation is centrally processed and analysed at the VŠB – Technical University of Ostrava[5]. This data has been handled and processed since the year 2000.

2.2 Reliability Computation

A majority of reliability computations is performed in the following way. The reliability computation of the whole system is executed on the basis of components reliability that are included in the system [6]. That is the reason why the reliability is computed in two phases. The first phase represents the retrieving of component reliability parameters and the second phase is the reliability computation itself. Other phases may include the evaluation of computed results and an improvement of the supply quality.

In virtue of experience, it is necessary to state that in most cases, the retrieving reliability parameter is far more complicated and involved than the reliability computation itself. The reliability rate is determined by an outage database by means of component reliability and global reliability indices.

Component Reliability Indices (System-Oriented Data) The majority of distributors create statistics for the component reliability of pieces of equipment such as breakers, lines, transformers, etc. They are especially collected to identify unreliable pieces of equipment and to be used as an input for probability computations of a systems behavior.

Global Reliability Indices (Customer-Oriented Data) This data is processed in the statistics oriented on consumer opinion concerning supply quality. The reliability of the supply to a consumer is usually assessed by the following so-called global indices. These indices are recommended for this purpose by UNIPEDE[6]. The meaning of these indices is the average reliability of a supply. They are primarily usable in relation to consulting companies, governing bodies and specific competitive analyses. Although, these indices are important in relation to common consumer requirements, we also need other indices for support of new consumer connections. In this case, we need indices related to the supply point.

2.3 Reliability Computation

Reliability is divided into two basic groups in compliance with the method of input data retrieval:

[5] `http://www.vsb.cz/`
[6] `http://www.eurelectric.org/`

– *Empirical reliability* - input data for the reliability computation is retrieved from data on equipment, or similar equipment operating under similar conditions
– *Predetermined reliability* - the probability of failure-free operation is expressed on the basis of knowledge about component status.

Obviously, incorrect input data leads to poor results, even when a correct computation method is applied. Moreover, in many cases of reliability computations in electrical power engineering, we face the problem of insufficient data size for a component, e.g. an insufficient number of historical records. In these cases, approximate input reliability data is often applied. Therefore, such reliability computation results serve further for a comparison of particular variants (e.g. network configuration); the same input reliability data being used for the reliability computation for each of the variants.

For statistically significant results of reliability computations, data on failures dating back to many years in the past is required. It is possible to compute basic reliability parameters of particular components from this database. These parameters are as follows: *annual failure rate*, *failure rate of a line*, *time to repair* ($MTTR$).

The number of failures per period is retrieved from the database. The period is usually defined depending on requirements concerning the reliability computation. An additional value necessary for the failure rate computation is the number of components for a set type and area. This value is possible to retrieve from the equipment owner. Other important information is possible to retrieve in more detailed databases, e.g. the most frequent cause of failures, areas of the greatest amounts of undelivered energy, etc.

3 Framework for Storage and Querying Outage Data

3.1 Outage Database

In [9], we have introduced a framework for storage and querying outage data [7, 6]. Since then, several works have been presented [8, 3, 5]. In this framework, databases of various distributors are transformed in a relation of the common relation scheme with 31 attributes (see Table 1). We see that some attributes are foreign keys of codebooks: these codebooks are labeled with an order number. Codebooks are often produced by an energy regulatory office, e.g. ERU[7] in the case of the Czech Republic. In this framework, we utilize a language for the transformation of distributor relations into the common relation. Once collections are stored in this relation, it is possible to query them with an SQL query. An architecture of the framework is depicted in Figure 3.

This framework consists of two parts. The first part addresses a storage of distributor relations, whereas the second part addresses querying the common database. We see that the framework includes the following input types:

[7] http://www.eru.cz/

Table 1. The `Outage` Relation Scheme

Order	Attribute	Data Type	Foreign Key/ Codebook Order
1	distributor	NUMBER	yes/01
2	event_order	CHAR	–
3	event_type	NUMBER	yes/02
4	distribution_point	NUMBER	yes/03
5	area	CHAR	–
6	network_type	NUMBER	yes/05
7	network_voltage	NUMBER	yes/04
8	equipment_voltage	NUMBER	yes/04
9	original_event_order	CHAR	–
10	event_cause	NUMBER	yes/06
11	equipment_type	NUMBER	yes/07
12	damaged_equipment	NUMBER	yes/08
13	damaged equipment_type	NUMBER	yes/10
14	amount	NUMBER	–
15	short_type	NUMBER	yes/09
16	producer	NUMBER	yes/11
17	production_date	DATE	
18	T0	DATE	–
19	T1	DATE	–
20	T2	DATE	–
21	T3	DATE	–
22	T4	DATE	–
23	TZ	DATE	–
24	P1	NUMBER	–
25	P2	NUMBER	–
26	D1	NUMBER	–
27	D2	NUMBER	–
28	Z1	NUMBER	–
29	Z2	NUMBER	–
30	LxT	NUMBER	–
31	failure_type	NUMBER	yes/13

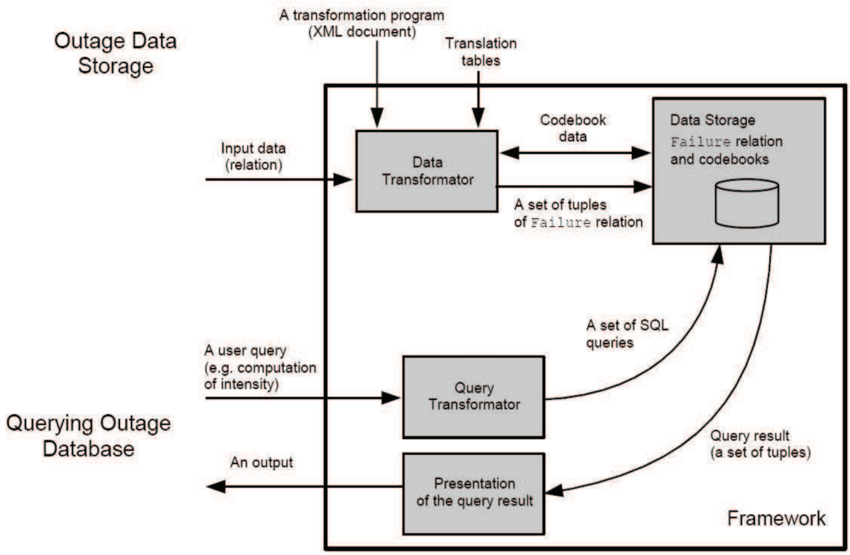

Fig. 1. Storage and Querying Outage Data Framework

1. *Input relation* – distributor's relation, often an xls file
2. *Transformation program* – a program transforms the input relation into the common relation
3. *Codebooks* – a codebook is a set of couples $\langle id, value\rangle$. For more efficient access, `id` attribute is often indexed. Codebooks are stored in data storage. Codebooks are often produced by an energy regulatory office, e.g. ERU[8] in the case of the Czech Republic.
4. *Translation tables* – tables translate terms of various distributors into common terms of codebooks. These tables are labelled with an order number. A translation table is a set of couples ⟨`term`, `translation` ⟩ used to unify terms. Translation tables are built by an author of the transformation program in cooperation with an expert of each distributor.

It may seem that it is necessary to build codebooks, transformation program, and translation tables for each data file. However, a distributor may deliver its data files per half year. Transformation programs and translation tables are sometimes the same for data files of the distributor. Consequently, the data processing is automated after development of these files related to the data transformation.

[8] `http://www.eru.cz/`

3.2 Data Transformation

Below, we describe the data transformation done by a transformation program coded as an XML document. Our own XML application, called *EcstlXML*, has been developed for this purpose. To completely describe the scheme would be beyond the scope of this paper (for detail see [8]), which is the reason why we only describe basic principles of this transformation.

Each tuple of the distributor's relation is read. A transformation program consists of a set of statements. Statements enable a transformation into the output value: transformation statements are processed for the output tuple's attribute and input tuple. In other words, one tuple from the output relation is built for each individual tuple of the input relation. The output relation is inserted into an RDBMS.

Transformation statements for one output value are implemented by one `Transformation` element in a transformation program. A set of these elements is the transformation program. This program must be manually implemented for each piece of distributor data. The example below shows some features of the transformation language.

Example 1. Retrieve year, month, day, hour, and minute of the outage in an input tuple in indexes 5, 6, 7, 8, and 9, respectively. Store this retrieved date into the output tuple in index 18 (T_0).

```
<Transformation OutputItemIndex="18">
  <Date>
    <Year>
      <InputItem TupleName="a" Index="5"/>
    </Year>
    <Month>
      <InputItem TupleName="a" Index="6"/>
    </Month>
    <Day>
      <InputItem TupleName="a" Index="7"/>
    </Day>
    <Hour>
      <InputItem TupleName="a" Index="8"/>
    </Hour>
    <Minute>
      <InputItem TupleName="a" Index="9"/>
    </Minute>
  </Date>
</Transformation>
```

It may seem that the number of these statements is rather low. In the case of distributors, we must define about 80 statements addressing many anomalies in the data.

4 Debugging Tool

4.1 Motivation

Initially, we must answer a question why it is necessary to utilize a debugger to tune transformation programs. There are primary two reasons. The first reason is that transformations can be relatively complex, for example see a transformation in Figure 2. Although it seems that we tune a transformation for all input data files, it is not true, because distributors provide their data for different schemes, moreover input data can be different among various years for one distributor. Therefore, we must write transformations for almost each input data file which is rather mistake-susceptible.

```
<Transformation OutputItemIndex="20">
  <Date>
    <FormattedDate FormattingString="H':'mm':'ss">
      <InputItem TupleName="a" Index="45"/>
    </FormattedDate>
    <FormattedDate FormattingString="d'.'M'.'yyyy"
             AlternativeFormattingString="M'/'d'/'yyyy">
  <InputItem TupleName="a" Index="44"/>
    </FormattedDate>
  </Date>
</Transformation>
```

Fig. 2. A transformation of the output attribute 20

The second reason depends on the first one. In the system, we must process a high number of various input files to transform outage data to the unified database. For example, in our case (outage data from the Czech Republic and Slovak Republic, see the first section), we must process the following files:

– approximately 150 transformation files
– approximately 150 data files
– approximately 550 translation and other files used

It means, we must handle 850 files to build the unified outage database including 309,000 records. Evidently, such a system is very mistake-susceptible. Since we need as many as possible correct records for correct reliability computation, we need a tool finding incorrect transformations or input files. One of this tool is a debugger.

4.2 Debugging Tool

Our debugger enables us to load a transformation program and all files used for the transformation and step each command of the transformation. For each command, input and output values are shown. In this way, we can distinguish incorrect transformations or other input files and fix them. Wrong transformations are distinguished by debugging one or few number of records from input data files. Therefore, it is not necessary to debug the whole, potentially big, data file. In Figure 3, we see a screenshot of this debugger.

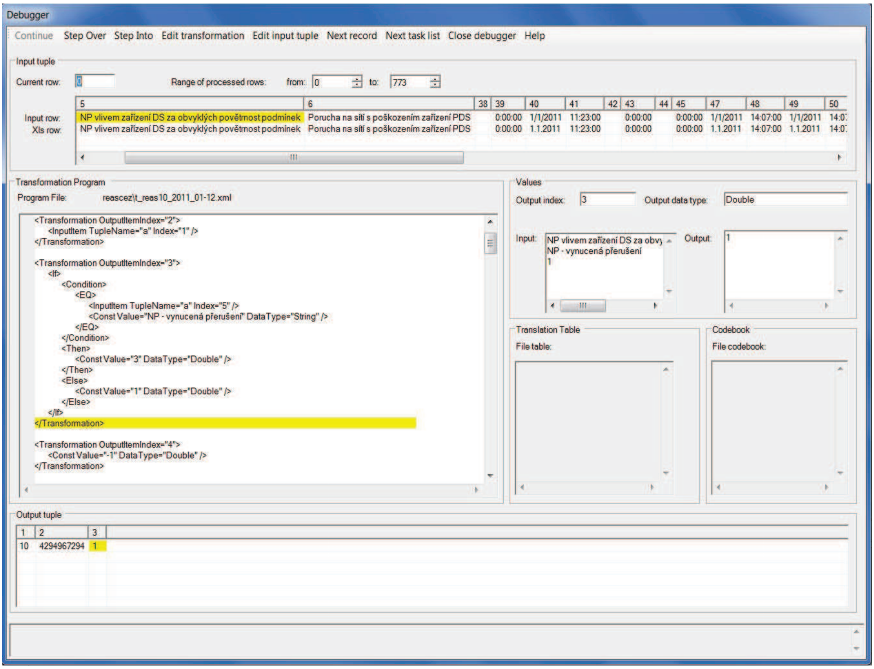

Fig. 3. Tool for debugging of data transformations

This tool is currently utilized in the following cases:

- Cleaning old transformation programs to find and fix incorrect transformations and files.
- Tuning transformation programs for new data.

In this way, we have found a number of incorrect transformations and fixed them. As the result, more correct transformations means the more correct outage database and it means more correct reliability computation.

5 Conclusion

A statistical significance of an outage database depends on the number of records in the database. A larger database would describe the real condition of network equipment more accurately. Therefore, it is necessary to merge databases of various distributors. The main problem of the fusion is the heterogeneity feature: databases of various distributors differ from one another. We introduced a framework which enables the retrieval of parameters from various databases. This framework does this by using the unification of various data sets. Since this framework utilizes automatic transformations of outage records, we need to provide a tool for debugging of transformations. In this paper, we described this tool and we showed that it enables us to minimize errors in transformations. In our future work we must develop other methods to validate automatic transformations.

References

1. L. Bain, M. Englehardt, and M. Engelhardt. *Statistical Analysis of Reliability and Life-testing Models: Theory and Methods.* Marcel Dekker Ltd., 1991.
2. R. E. Barlow and F. Proschan. *Statistical Theory of Reliability and Life Testing: Probability Models.* Holt, Rinehart and Winston, Inc., 1975.
3. R. Bača, M. Krátký, and V. Snášel. Bulk-loading of Compressed R-tree with Failure Data. In *Proceedings of the 4th Workshop ELNET 2007.* FEECS, VŠB–Technical University of Ostrava, 2007.
4. W. H. Beyer. *Crc Standard Mathematical Tables.* CRC Press, Boca Raton, USA, 27th edition, 1984.
5. P. Chovanec and M. Krátký. Benchmarking of Lossless R-tree Compression for Data of Failures in Electrical Power Networks. In *Proceedings of the 7th Workshop of ELNET, Czech Republic*, 2010.
6. R. Goňo and S. Rusek. Analysis of Power Outages in the Distribution Networks. In *Proceedings of the 8th International Conference on ElectricalPower Quality and Utilisation, EPQU 2003. Cracow, Poland.* IEEE Press, 2003.
7. R. Goňo, S. Rusek, and M. Krátký. Reliability analysis of distribution networks. In *Proceedings of the 9th International Conference on Electrical Power Quality and Utilisation, EPQU 2007. Barcelona, Spain.* IEEE Press, 2007.
8. M. Krátký, R. Goňo, and S. Rusek. A Framework for Querying and Indexing Electrical Failure Data. In *Proceedings of ELNET 2006. Ostrava, Czech Republic*, 2006.
9. M. Krátký, R. Goňo, S. Rusek, and J. Dvorský. A Framework for an Analysis of Failures Data in Electrical Power Networks. In *Proceedings of ELNET/PEA 2006. Gaborone, Botswana.* ACTA Press/IASTED, 2006.
10. N. R. Mann, R. E. Schafer, and N. D. Singpurwalla. *Methods for Statistical Analysis of Reliability and Life Data.* John Wiley & Sons, Inc., NewYork, 1974.
11. J. Piskač and J. Marko. Regulations for electric power system No.2 Failure statistics at electricity distribution, 1974.