DETECTION OF INCOHERENCES IN A TECHNICAL AND NORMATIVE DOCUMENT CORPUS

Susana Martin-Toral[1], Gregorio I. Sainz-Palmero[1,2]
[1]*Computer and Information Technologies Division. Fundación CARTIF*
Parque Tecnológico de Boecillo, 47151 Valladolid, Spain

[2]*Department of Systems Engineering and Control, School of Industrial Engineering*
University of Valladolid, 47011 Valladolid, Spain

Yannis Dimitriadis
GSIC - Group of Intelligent and Cooperative Systems, School of Telecommunications Engineering
University of Valladolid, 47011 Valladolid, Spain

Keywords: Document corpus, content incoherence, natural language processing, text mining, artificial intelligence, document engineering.

Abstract: This paper is focused on the problems and effects generated by the use of a document corpus with mistakes, content incoherences amongst its connected documents and other errors. The problem introduced in this paper is very relevant in any area of human activity when this corpus is used as base element in the relationships between company partners, legal support, etc., and the way in which these incoherences can be detected. These problems can appear in several ways, and the produced effects are different, but a common situation exists in those areas of activity where many linked documents must be generated, managed and updated by different authors. This paper describes some examples of this problem in the case of a technical document corpus used amongst partners, and the solution framework developed for this case. Several types of incoherence have been detected and formulated, connected with problems described in other research areas such as information extraction and retrieval, text mining, document interpretation and others, but all of them have been bounded and introduced from the point of view of document incoherences and their effects, specially in a company context. Finally the computational architecture and methodology uses are described and some initial results of incoherence detection are discussed.

1 INTRODUCTION

Documentation, on paper or in electronic format, is a base element for the information society. It is the most usual way to store, save and exchange information in a wide range of human activity contexts, so the information and knowledge contained in it has to be right and clear with no possibility of confusion or contradiction. But this goal is not trivial due to several facts. Some public and private sectors handle documentation that is not-methodologically generated, suffers changes and grows in volume and versions.

It is difficult to find organizations working with heterogeneous sets of connected documents that manage this movement in a suitable and formal way, with a unique formulation in their generation, management and control (see Figure 1), so the problem of incoherences in related documentation appears: mistakes in the cross references, redundant, contradictory, missing or wrong information.

The case involved in this paper comes from the use in technical documentation by a company of the electrical sector. In this context, an "incoherence" represents any lack of consistency amongst related documents or even within the same document.

For example, some types of incoherences appear when a document does not match the correct structure, according to the known rules for its generation. In this situation a structural incoherence appears. Other types of incoherence could be numerical. This concerns the numerical values contained in a document that must agree with the values indicated in the norm, standard or reference document. A contradiction in numerical values between documents with the

Martin-Toral S., I. Sainz-Palmero G. and Dimitriadis Y. (2008).
DETECTION OF INCOHERENCES IN A TECHNICAL AND NORMATIVE DOCUMENT CORPUS.
In *Proceedings of the Tenth International Conference on Enterprise Information Systems - AIDSS*, pages 282-287
DOI: 10.5220/0001699102820287

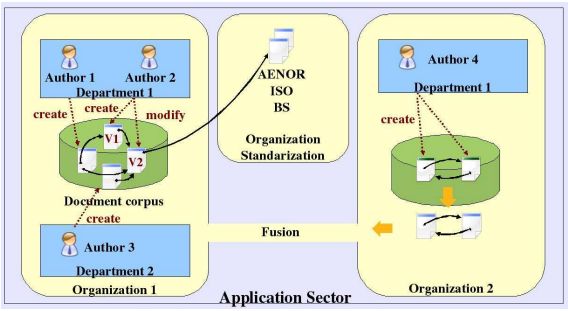

Figure 1: Documentation generation scheme in a technical business.

same concept is not possible. These two examples, and other types of incoherences, makes up an initial definition of an incoherence taxonomy.

The impact of all these problems for an organization, both in its internal and external relationships, could cause economic, legal, technical, even serious social consequences, so when this happens there is a great interest in eliminating them. Thus, some sectors with economic activity have been seen to show a growing interest in solving this kind of problem, though of course, such interest is not exclusive to such sectors.

- *Civil Engineering & Electrical sector*, in general any industrial environment in which collaborative partners work with related technical documents, normatives and standards (CARTIF, 2006).
- *Software industry*, in which the documentation generated from a unified software process is affected by the problems described previously (Arango, 2003).
- *Healthcare sector*, the volume of documentation managed is enormous and very critical due to the effects of mistakes in such documentation (Mingshan and Ching-to, 2002).
- *Legal and Law sector*, when laws or legal norms are contradictory, "legal antinomia" (Ruiz, 2002).

Otherwise, three different parts could be affected by these problems:

- *Owner*, economic and legal aspects: incoherent documentation does not match the application norms and policies to the sector, which could result in third party damages.
- *User*, by employing a contradictory, ambiguous or even wrong documentation with similar effects to those commented previously.
- *Consumer*, using products and services created by weak and incoherent documentation that can cause bad quality or dangerous products or services.

At this point, the interest concerning the detection and elimination of these incoherences appears. Documentation free of incoherences could solve a relevant problem, or at least that focusing on wrong or confused information, facilitating a coherent management of that documentation and obtaining a better quality of products and services. This is the main motivation for dealing with the problem.

The organization of the rest of the paper is as follows: first of all a definition and classification of the detected incoherences in the case involved, and techniques that could be applied for its detection, are presented. In the technological framework section the computational and conceptual architectures developed in this work are described. Finally, the most interesting results obtained are discussed and the main conclusions of this work are put forward.

2 DOCUMENT INCOHERENCES. AN APPROACH

Once the problem and the motivations have been presented, the next step is to formally define what is considered an incoherence in this work, taking into account the documentation involved (see (CARTIF, 2006)): *content incoherence is seen as the weakness of consistency amongst related documents, or amongst different pieces of the same document, or the lack or excess of information in a document.*

This definition introduces subjectivity in deciding what can be considered as an incoherence and its effects, thus its importance. To assess this issue, an initial incoherence taxonomy has been defined:

- *Structural*, concerning the logical rules for document generation: differences in style or format employed when the document was generated or updated. This aspect is connected with research into document analysis and the logical and layout document structure (O'Gorman and Kasturi, 1995).
- *References*, documents use references to other documents, norms or standards in order to support the document content or to avoid describing any aspect explained in the references. The incoherence could happen when the reference is not adequate, or does not exist, or is not referenced.
- *Numerical*, this concerns the numerical values contained in a document that must agree with the values indicated in the norm, standard or document of reference. A contradiction between documents for the same concept is not possible.

- *Measure Units*, the units for measuring used in a document have to work correctly in accordance with the standard or the International System of Units.
- *Attribute*, which is similar to the numerical one, but applied to attributes such as colors (green, black, red), shapes (square, triangular), states (liquid, solid, gaseous), etc.
- *Denomination or conceptual*, it is very important to use the concepts in a suitable way for the context involved. A conceptual incoherence could happen if an important concept is denominated of different ways in the same document or even in different documents.
- *Update*, the new version could contain more or less information or contradictory information with regard to the previous version.
- *Titles or subtitles* of a document do not match the contents of their sections.
- *Dirty words*, use of badly written words or words that do not exist.

In the technical context involved, each of these incoherences has a different relevance and effect, which is usually defined by the domain expert. Generally, depending on the type of incoherence to be detected automatically, it will be necessary to apply different techniques for information processing.

3 TECHNOLOGICAL FRAMEWORK PROPOSAL

After the theoretical introduction to the problem, this section deals with an approach to the technological framework for incoherence detection in documents. For this aim, a computational architecture has been proposed (see Figure 2) introducing different levels of information processing.

- *Preprocessing* converts the original documentation to an open format that facilitates its processing. For this task, the standard OASIS Open Document Format (ODF) has been selected (OASIS, 2007).
- In *extraction techniques* module, text and document mining techniques are mainly applied to extract the adequate information from each document in order to detect the several types of incoherence. Different representations of the same document must be used to detect the types of incoherence considered.
- *Representation of atomic information* keeps the suitable representations of the documents. Several types of document representation are needed according to the relevant type of contents or concepts to be checked.
- *Document representations by incoherence*. Here the most adequate representations of a document are taken from the previous module, according to the incoherence to be detected. Several representations could be required for an incoherence.
- *Comparison techniques* module uses text document mining techniques for matching different document representations to detect similarities and differences between them. This information will be used as a source of incoherence problems.

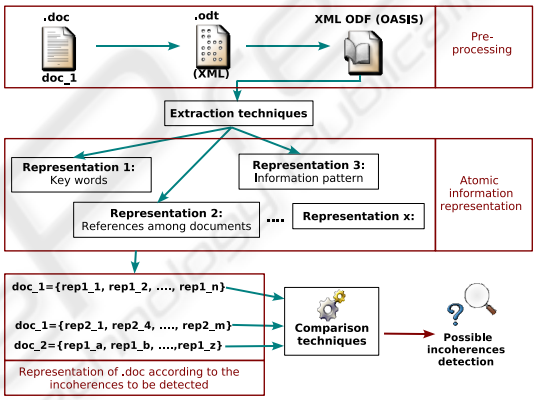

Figure 2: Computational architecture.

Once any potential incoherence has been detected and classified, it is the domain expert who must take the final decision about the relevance of the detected problem. Accordingly, the system could be improved by an adequate feedback.

In this work different techniques have been used to cover the functionalities of extraction and comparison modules. Most of them, mainly for extraction techniques, are based on the use of heuristic solutions, with similar criteria as to (Krulwich and Burkey, 1997).

3.1 Extraction Techniques

The adequate document representations have been obtained by different techniques of information extraction (Nahm, 2004), information retrieval (Berry, 2004), and document description (Jain et al., 2000). The proposal, from this point of view, could be seen as a pattern recognition problem, where every document is a pattern of the problem space and each document representation is the result of a feature selection

process, according to the knowledge of the document corpus. In this way, three main document representations (conceptual models) have been used:

1. *Key term representation*, using the Vector Space Model (VSM) (Salton et al., 1975). This type of representation facilitates the detection of reference, conceptual or title incoherences.
2. *Reference representation*. Every document is described according to the references to other documents. This representation mainly permits the detection of reference incoherences.
3. *Representation using relevant information patterns*. Here the information extraction is based on heuristics, according to information patterns detected inside the document corpora that are relevant in the domain. An example of this is the representation of a document by its technical data terms. Each one is represented by an "N-tuple", here $N = 4$:

   $< Term\ ;\ Operator\ ;\ Value\ ;\ Units >$

   Where *Term* is the word, or set of words, representing a relevant concept, Operator indicates that a term is bigger (>), smaller (<), than or equal to (=) a specific value, Value represents the numerical value, or enumerated data (colour, state, shape) of the term, and finally, Units is only used when the value is numerical and with units. Then the document is summarized by a set of this type of N-tuple. These N-tuples have been generated by Episode Rule Mining techniques (ERM) (Mannila et al., 1997). An example of a real 4-tuple is:
   $< wire\ CCX\text{-}56\text{-}D\ section\ ;\ =\ ;\ 54{,}6\ ;\ mm^2 >$
   This representation facilitates the detection of numerical, measure and attribute incoherences, applying suitable matching techniques.

3.2 Comparison Techniques

Matching techniques must be applied to obtain similarities, differences, deviations, and trends amongst document representations. Here, different text mining techniques have been used:

- *Classification* (Berry, 2004), based on VSM representation and Naive Bayes, KNN (K Nearest Neighbour) or TFIDF (Term Frequency - Inverse Document Frequency) classifiers, and using the libbow library (McCallum, 1996).
- *Clustering*, using VSM and a Hierarchical Expectation-Maximization (HEM) algorithm (Jain et al., 2000).
- *Summarization* using MEAD, a public domain portable multi-document summarization system (Otterbacher et al., 2002).
- *Trend detection* (Berry, 2004), using the edit distance, cosine similarity, and summarization techniques (Mani and Bloedorn, 1999).

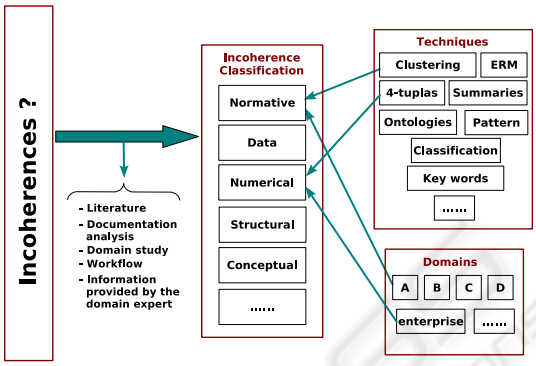

Figure 3: Conceptual architecture.

In the bibliography, some of these mentioned techniques are used to deal with problems not related to content incoherence detection. For this reason, the potential techniques discovered in the bibliography have been adapted to solve the incoherence problem, as in Figure 3.

4 EXPERIMENTS AND RESULTS

The experimental work has used the more suitable document representations and comparison techniques to detect potential incoherences. The effects and relevance of each one has been decided by the domain expert.

The document collection used in the experimental part consists of 873 documents corresponding to technical manuals (469 TM) and norms (404 N) of an electrical company. The documentation contains semistructured and connected documents, with many numerical Tables and Figures. It has been generated by multiple authors and branches, so the document corpus presents incoherence problems.

The experiments developed have covered two types of analysis, according to the principles of document engineering: structural and content analysis.

Structural analysis has been used to detect structural incoherences in the documentation: lack of mandatory sections, badly numbered chapters, etc. The knowledge of the rules about standard generation of documents permits us to process the document corpus. The results of this experiment have discovered that almost 30% of the norms present some type of structural incoherence, so they do not comply with the defined rules. For example, all the norms must have a section indicating what other norms are referred to

in the rest of the document. In this study, a 15.84% of the norms do not match with the rule because of the lack of some referred norm (see Table 1).

Table 1: Partial results for strutural incoherences.

Mandatory sections	Coh.norms	Incoh.norms
Introduction	134 (33.17%)	270 (66.83%)
Referred norms	340 (84.16%)	64 (15.84%)
Application field	365 (90.35%)	39 (9.65%)

In content analysis, the main objective is to detect possible incoherences by processing the documentation contents, but not its structure. For that end, document representations shown in section 3.1, and comparison techniques described in section 3.2 have been used.

The use of the Vector Space Model allows us to discover relevant terms that should not present conceptual incoherences. Using this representation, terms written in different languages can be detected, and they are also considered as incoherences by the domain expert.

With classification experiments, the trend of document clustering can be studied to detect strange behaviours. In the results, we can see how some documents tend to be classified in a non adequate class or category. The best results, see Table 2, have been obtained with the TFIDF classifier.

Table 2: TFIDF classification results for norms.

Trial	Ok classified/total	% Accuracy
0	92/121	76.03
1	99/121	81.82
2	94/121	77.69
3	95/121	78.51
4	101/121	83.47
5	94/121	77.69
6	90/121	74.38
7	95/121	78.51
8	91/121	75.21
9	89/121	73.55
	Average	77.69 stderr 0.94

From the total available norms, a 70% has been selected to train the classifier, and the rest (121 documents) to make the test. The average accuracy for 10 trials is 77.69%.

Both document classification and clustering, are developed according to relevant terms of each document, so incoherence traces can be detected by analyzing the content of these bad classified documents, in comparison with the rest of the members of the category. Similar situations appear when a document is clustered in an unsuitable cluster. These situations are potential sources of incoherences. All the norms are coded according to the material family they represent, so this codification can be used as a natural initial classification. When a norm is classified or clustered in a group not belonging to its material family, according to the codification, this is because its content is more similar to the other category, so it presents similarities to an other material family. The analysis of this strange behaviour can report traces of incoherence problems. For example, the strange situation appears when the norm N 72.30.03, that belongs to the transformer family (code 72), is classified into the wire family (code 56) because of its content (in the electrical sector).

Representing the documentation using the reference model facilitates the detection of reference incoherences. The representation is in the following form:

```
TM1  N1  N3
TM2  N1  N4  N5  N9
TM3  N2  N5
....
TMN  N2
```

where TM are technical manuals and N are norms. With this information, unsuitable use of documents can be detected, and also incoherences related to the use of non-existent and wrong coded norms. It has been detected that a 5% of the referenced norms present a wrong codification, and more than a 7% are non-existent, maybe because they are deprecated and have been eliminated from the final version of the document corpus.

Finally, the representation using 4-tuples allows the detection of numerical, measures and attribute incoherences. The experimental part has been applied to obtain all 4-tuples from the documentation, applying ERM techniques to extract this information pattern. Matching the representations between two related documents, or matching the N-tuples of the same documents, facilitates the detection of this kind of incoherences. They seem to be the more numerous incoherences, so there is an increasing interest in their detection. Thus, at the moment, a detailed study of this type of representation, and all the related techniques, is being considered.

5 CONCLUSIONS

This paper introduces the problem of document incoherences from the point of view of the organizations, companies or environments using document corpus with mistakes, wrong documents and confused contents, and the effects of this inadequate documentation: economic, legal, technical and social.

Taking into account the technical and normative documentation involved in this work, an attempt to define and classify incoherence has been introduced, which could be used in most technical contexts. From that, each incoherence has been connected with several research areas (information extraction and retrieval, document analysis, etc.) in order to find the best way to detect the incoherence by information processing. The results obtained by experiments have allowed us discover several categories of incoherences, even some unknown to the domain experts.

The study of new domain and sector documentation could expand and improve the proposed incoherence classification, but not all incoherences have the same relevance or the same importance for the affected sectors. The experimentation in this work has tried to apply the more suitable techniques to detect those with the most relevant impact in the affected areas.

To achieve all these objectives, different document representations and comparison techniques have been applied. In this aspect, a new relevant information pattern, repeated in technical documentation, has been used (N-tuples), allowing the detection of one of the most important and negative incoherence types found in technical domains: numerical incoherences.

The interpretation and evaluation of the results have been developed in both unsupervised and supervised ways, in this latter case, with the help of the domain expert. From this evaluation, different levels of incoherence have been detected:

- Some experimental results have shown strange behaviours, and therefore the presence of potential incoherences. A deeper study could be needed to detect specific incoherences. These results are obtained by classification or clustering methods.
- Other results have directly shown potential cases of incoherences, and the help of the domain expert is only needed to ensure that the problem exists. This is the case of incoherences of wrongly coded and non-existent norms, structural, or content incoherences using VSM, and numerical measures and attribute incoherences using 4-tuples.

Due to the existence of incoherence and its negative effects, for both to organization and citizens, future work could deal with the definition of a new methodology for the generation of new documentation free of incoherences, to avoid the initial seed of the problem.

ACKNOWLEDGEMENTS

This work has been supported in part by the Spanish Industry, Tourism, and Commerce Ministry through the project FIT-350100-2006-272.

REFERENCES

Arango, F. (2003). *Gestion de inconsistencias en la evolucion e interoperacion de los esquemas conceptuales OO, en el marco formal de OASIS*. PhD thesis.

Berry, M. W. (2004). *Survey of Text Mining : Clustering, Classification, and Retrieval*. Springer.

CARTIF, F. (2006). Gestor documental de normativa (DOCNOR). PROFIT project (PROgrama de Fomento de la Investigacin Tecnologica). Project reference: FIT-350100-2006-272.

Jain, A. K., Duin, R. P. W., and Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(1):4–37.

Krulwich, B. and Burkey, C. (1997). The infofinder agent: Learning user interests through heuristic phrase extraction. *IEEE Expert: Intelligent Systems and Their Applications*, 12(5):22–27.

Mani, I. and Bloedorn, E. (1999). Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1-2):35–67.

Mannila, H., Toivonen, H., and Verkamo, A. I. (1997). Discovery of frequent episodes in event sequences. *Data Min. Knowl. Discov.*, 1(3):259–289.

McCallum, A. K. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow.

Mingshan, L. and Ching-to, A. M. (2002). Consistency in performance evaluation reports and medical records. *The Journal of Mental Health Policy and Economics*, 5(4):191–192.

Nahm, U. Y. (2004). *Text mining with information extraction*. PhD thesis. Supervisor-Raymond J. Mooney.

OASIS (2007). OASIS. *Organization for the Advancement of Structured Information Standards*. URL: http://www.oasis-open.org. Last visit: February 2007.

O'Gorman, L. and Kasturi, R. (1995). *Document Image Analysis*. IEEE Computer Society, Los Alamitos, California, USA.

Otterbacher, J., Radev, D., and Luo, A. (2002). Revisions that improve cohesion in multi-document summaries: A preliminary study. In *Proc. of the Workshop on Automatic Summarization (including DUC 2002)*, pages 27–36. Association for Computational Linguistics.

Ruiz, M. (2002). *Sistemas jurdicos y conflictos normativos*. Dykinson, Universidad Carlos III de Madrid, Instituto de Derechos Humanos Bartolom de las Casas.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communication of the ACM*, 18(11):613–620.