2009 10th International Conference on Document Analysis and Recognition

# Detection of incoherences in a document corpus based on the application of a neuro-fuzzy system

S. Martin, V. Arribas
Fundación CARTIF (*)
Parque Tecnólogico de Boecillo
47151 Valladolid, Spain
susmar@cartif.es, vicarr@cartif.es

G. I. Sainz (*)
Dept. of Systems Engineering and Control
School of Industrial Engineering
University of Valladolid, Spain
gresai@eis.uva.es, gresai@cartif.es

## Abstract

*The aim of this paper is to detect incoherences in concepts, ideas, values, and others contained in technical document corpora. The way in which document collections are generated, modified or updated generates problems and mistakes in the information coherency, leading to legal, economic and social problems. A solution based on summarization, matching and neuro-fuzzy systems is proposed to dealt with this problem.*

*For this goal, every document (from the electric domain) is summarized by its relevant information in the form of 4-tuples of terms, describing the most relevant ideas and concepts that must be free of incoherences. These representations are then matched using several well-known algorithms (Levenshtein distance and cosine similarity). The final decision about the real existence or not of an incoherence, and its relevancy, is obtained by training a neuro-fuzzy system FasArt in a supervised classification process, based on the previous knowledge of the activity area and domain experts.*

*On the other hand, using this fuzzy approach, it is possible to extract the learnt and expert knowledge from the the neuro-fuzzy system, through a set of fuzzy rules that can support a decision taking system about this complex and non objective problem.*

## 1. Introduction

Documents, on paper or in electronic format, are base element for the society's activities. 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 very 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, so the problem of incoherences in related documentation appears: mistakes in the cross references, redundant, contradictory, missing or wrong information, or, in general, rules for quality documentation are not achieved [8].

The impact of all these problems in 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 detecting and eliminating them. Thus, some sectors show a growing interest in solving this kind of problem: healthcare services [9], software companies, the legal and law sector, civil engineering [8], etc.

Documentation free of incoherences improves a coherent management of it and a better quality of the products and services generated. But when any solution aims to deal with this problem, other important difficulties appear: How/What is a document incoherence? Does every incoherence have the same relevancy? In both cases the answer is subjective and depends on the industrial and economic sector and the know-how of the domain experts.

This paper deals with incoherences in documents by summarizing and matching techniques, then the expert and subjective knowledge is incorporated by a supervised learning based on the neuro-fuzzy system FasArt. In this way, it is possible to detect incoherences in documents, so their classification and the learnt knowledge can be extracted by fuzzy rules, explaining the way in which the expert takes his decisions about the case. These fuzzy rules can be a support for a decision taking system.

The organization of the rest of the paper is as follows: first of all, a tentative definition and classification of the detected incoherences in the case involved are presented,

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.101

along with techniques that could be applied for its detection. Next, the proposal of this paper to deal with document incoherences is introduced, describing its several phases. Then, the experimental procedure done to test the proposal is shown in Section 4. Finally, the most interesting results obtained are discussed and the main conclusions of this work are put forward.

## 2. Incoherence in documents

The approach introduced in this paper combines general concepts and techniques with heuristics about incoherences and their contexts. This can, in general, be a very practical approach, due to the difficulties in defining when an incoherency appears in a document and its importance, which depends greatly on the domain and experts.

At this point it is necessary to give some type of description of what is considered to be an incoherence in this work: *an 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 it* [8].

This description introduces subjectivity about what can be considered an incoherence and its effects/relevancy, thus its importance. From the document collection involved in this work, about the electric domain, some interesting types of incoherences cause negative effects in this domain, in accordance with the domain experts:

- *Numerical and Attribute* incoherences concern the numerical values and technical attributes (such as colours, shapes, states, etc.) 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.
- *Conceptual* incoherences happen when an important concept is denominated in different ways in the same document, or even in different ones. It is very important to use concepts in a suitable way for the context involved.
- *Reference* incoherences happen when documents use references to other documents, norms or standards, to support the document content or to avoid describing any aspect explained in the references. The incoherence appears when this reference is not adequate, does not exist, or is not referenced.

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

In technical and scientific literature, the formulation of the problem involved in this paper is not very usual, at least with the same meaning, but there are well-known techniques that can be applied in the detection of document incoherences: text mining, pattern recognition, semantic analysis, etc. In general, most of them, mainly for extraction techniques, are based on the use of heuristic solutions, with similar criteria as in [6]. In [3] the discourse coherence using Latent Semantic Analysis (LSA) is described, recognizing that it is not usual to translate the general and theorist ideas about coherence into a practical level.

This work is focused on the summarization of a document corpus using 4-tuples for the detection of numerical incoherences. This kind of incoherence is very damaging in the electrical domain, with high economical consequences.

## 3. Detection and classification of incoherences: an approach

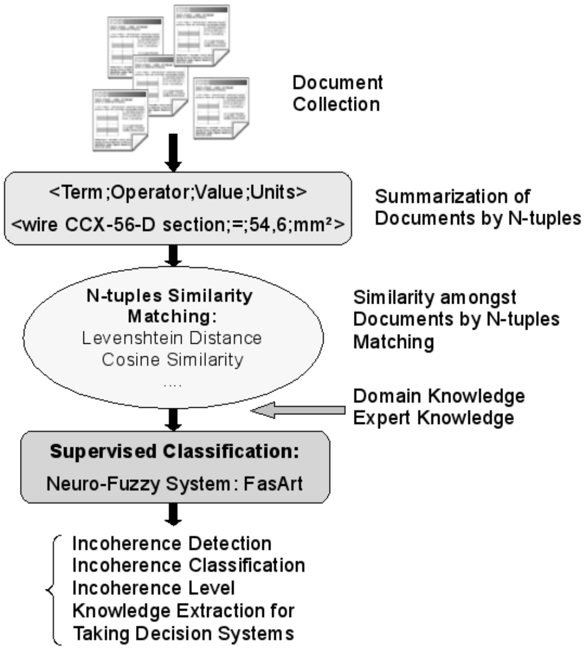
The main goal to be reached in this approach is to detect when an inconsistency is contained in a document. With this aim, the procedure shown in Figure 1 is carried out: the documents involved are summarized by a set of key terms and concepts that are very relevant in the domain, and in accordance with the incoherence types described in section 2. In this way, documents are summarized by a set of N-tuples (see section 3.1). At the moment, this is a semi-automatic procedure based on extraction techniques.

The next step focuses on the use of matching techniques to establish the level of similarity between the elements of every two N-tuples, to decide, in a subsequent step, if there are incoherences in the document contents or amongst documents. Here, well-known techniques, such as the Levenshtein distance [2] or the Cosine similarity [1], are used.

At this point, a critical aspect is to decide when two document pieces are incoherent, or even the incoherency degree. This decision concerns the experts of the document domain in most of the cases. This aspect is approached by a supervised learning in which the knowledge of the expert is taken into account. Here a neuro-fuzzy system based on FasArt [10] is used. Although other solutions could be used, this type of systems have been used in previous works for pattern recognition and knowledge extraction [11] with reasonable results.

The final goal obtained is the detection, and classification, of incoherences amongst document pieces. An inconsistency degree for each case is provided by the fuzzy approach. On the other hand, it is possible to generate a further result using this approach: a knowledge base using fuzzy rules about the way in which incoherences are detected, that will be used, for example, to generate a free incoherences editor for technical documents.

**Figure 1. Approach based on neuro-fuzzy system for detection of document incoherences .**


Document Collection
<Term ;Operator;Value ;Units>
<wire CCX-56-D section;=;54,6;mm²>
Summarization of Documents by N-tuples
N-tuples Similarity Matching:
Levenshtein Distance
Cosine Similarity
....
Similarity amongst Documents by N-tuples Matching
Domain Knowledge
Expert Knowledge
Supervised Classification:
Neuro-Fuzzy System: FasArt
Incoherence Detection
Incoherence Classification
Incoherence Level
Knowledge Extraction for Taking Decision Systems

## 3.1. Summarizing documents by 4-tuples

In this work, information extraction techniques are used to obtain representations that summarize each document of the corpus. Here, the information extraction is based on heuristics [6], according to the information patterns detected inside the document corpora that are relevant for the experts in the electrical domain. An example of this is the summarization of a document by its technical data terms. Each one is represented by an "N-tuple", here $N = 4$:

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

$<$*Term ; Operator ; Attribute ;* $>$

Where *Term* is the word, or set of words, representing a relevant concept, *Operator* can indicate that a term is bigger, smaller than, or equal to a specific value/attribute, *Value/Attribute* represents the numerical value, or an attribute (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 similar approaches to Episode Rule Mining techniques (ERM) [7]. An example of real 4-tuples are:

$<$ *wire CCX-56-D section ; = ; 54,6 ;* $mm^2$ $>$

$<$ *cover of wire CCX-56-D colour ; = ; green ;* $>$

This representation facilitates the detection of numerical, measure and attribute incoherences, applying suitable matching techniques, such as those used in this work, by their relevance in the domain involved in this paper. When two 4-tuples present the same information in their four elements, it is certain that there is not incoherence in this information. If two 4-tuples present the same information in all their elements but different values, then a numerical incoherence exists. In the rest of the situations, the domain and expert knowledge is needed to define the existence or not of this kind of incoherence and its relevance, and similarity measures are used to technically define every situation. The same methodology is applied for measure and attribute incoherences.

## 3.2. Similarity measures for tuple-elements

Two approaches have been considered for similarity measures amongst n-tuple elements: based on edition distance and vector space. The first group is based on how many changes and which type of changes are necessary for turning a character string into another one. Three main operations are identified within this topic: insertion, deletion and substitution. The relevance of each one is tuned by the user. Within this group of measures, the following can be found: Levenshtein distance [2, 1] and Needleman distance [1].

The result is zero whenever two strings are identical. If differences exist, the distance is an integer number greater than zero.

The second group is oriented in token-based distances, which computes distances between two groups of words (tokens). Within this group, the following can be found: *Cosine similarity* [4, 1], *Jaccard similarity* [2, 1], *Dice coefficient* and *Overlap coefficient* [1].

In this work, documents contain tuples made of four terms (see section 3.1). Taking this into account, cosine similarity has been proposed to establish term similarities. This approach is also supported in works such as [5] and [2], where different methods for string matching are evaluated in other contexts. On the other hand, operators, numeric values and units are short-length strings of characters of one-word size, thus being better to apply edition distances, as what is to be measured is the difference between two of them.

The aim of comparing two tuples is to detect the existence of incoherences amongst the contents expressed in them. These results are the input for the neuro-fuzzy system, FasArt. Expert and domain knowledge is needed along with similarity measures for generating a supervised system incorporating this expert but no objective knowledge.

## 3.3. Neuro fuzzy system FasArt

The FasArt model [10] is a neuro-fuzzy system based on the Adaptive Resonance Theory (ART): Fuzzy ARTMAP.

FasArt introduces an equivalence between the activation function of each FasArt neuron and a membership function. In this way, FasArt is equivalent to a Mamdani fuzzy rule-based system with: Fuzzification by single point, Inference by product, Defuzzification by average of fuzzy set centers. A full description of this model can be found in [10].

The FasArt system has been used in several previous works [11, 12] for modeling, fault detection, pattern recognition, etc. with reasonable results when its accuracy as a fuzzy model is involved. Knowledge extraction of the knowledge learnt can be done using this neurofuzzy system by a set of fuzzy rules.

## 4. Experimental Methodology

Documents involved in this work have been summarized in 4-tuples by a semiautomatic procedure. Two sets of documents containing 4-tuples were generated as follows:

1. Document corpus from a power company of the electric domain containing standards, protocols and operating manuals about usual tasks to be carried out by the company and partners. This collection consists of 10 documents that were summarized by N-tuples with the most relevant terms or concepts from the point of view of the company experts. A total of 40 tuples are within this group.

2. Synthetic document set of 4 documents generated by a manual procedure holding 29 ideal tuples, along with consistent and inconsistent variants. The consistent ones having slight changes in the term: inclusion of stop words, words beginning in upper or lower case, etc, a total of 234.

Once experimental data is ready, matching techniques have been applied to calculate similarity measures in the way described in section 3.2. The matching has been done element by element of the tuple, a global similarity measure being calculated as the average of element similarities. Each case has been evaluated by a domain expert, who decides whether if there is incoherence or not.

The neuro-fuzzy FasArt system has been used in this case to learn this knowledge about incoherences contained in the 4-tuples documents. The system output is the presence of incoherence regarding the similarity measures between two tuples as its input. Two alternatives were considered as inputs: 4 similarity values (one for each term tuple), and 5 similarity values (here an extra similarity value about the whole 4-tuple is considered). The FasArt system has been tuned with respect to vigilance factor $\rho$ (ranging from 0.1 to 1.0) and fuzzification rate $\gamma$ (ranging from 1.0 to 20.0).

On the other hand, the system has been trained and tested by cross validation: a) Training and testing using synthetic documents; b) Training and testing using real documents; c) Training with synthetic and testing with real documents; d) Training with random groups of 70% of all tuples and testing with the remaining 30%.

Each experimentation alternative has been evaluated by analyzing the detection error and complexity of the system, through the number of fuzzy rules from the neurofuzzy system. This aspect is very relevant because this knowledge base could be used to generate a decision-taking system about document incoherences, i.e, a free-incoherences document editor.

## 5. Results and analysis

Attending to the error rate, most stable and coherent results were achieved by training and testing the system with random groups of synthetic and real tuples similarities (see Table 1). Although this experimentation presents a more realistic component (in comparison with the use of synthetic tuples), it does not offer the best classification results.

**Table 1. Classification using heterogenous training sets.**

| Inputs | Similarity | Complexity (rules) | Error |
|---|---|---|---|
| 4 | Lev. | 66 | 4.9% |
| 5 | Lev. | 97 | 3.4% |
| 4 | Lev. | 201 | 5.0% |
| 5 | Lev. | 242 | 4.7% |
| 4 | Need. | 100 | 4.8% |
| 5 | Need. | 68 | 4.9% |
| 4 | Need. | 435 | 4.5% |
| 5 | Need. | 506 | 4.6% |

In this case, the best results are achieved when the Levenshtein distance and whole similarity value about the 4-tuple are considered: detection classification is 3.4%, and the number of fuzzy rules is 97, which is a moderate complexity. When no global similarity is involved, the error is 4.9%, but the resultant system is simpler than in the previous case, with 66 fuzzy rules. In the other alternatives involved, the results are very similar to the previous ones, with an error near 5%, but the complexity reached is higher, specially when the Needleman distance is used.

On the other hand, the best rate of classification is obtained when the system is trained with synthetic and tested with real tuples ($error = 0.1\%$ ), followed by an error of 0.5% when synthetic tuples are used throughout the process. Both results show that synthetic tuples cover all the incoherence cases held within the set of real ones, as the system is able to recognize almost all of these latter ones. But a more realistic scenario should include real tuples for the training stage. Finally, training and testing the system with real tuples shows the worst case in the experimentation, which has an error of $63.2\%$, which is feasible, as they do not cover the same cases as the synthetic ones.

Therefore, the use of random groups validates the experimental stage, as a success rate of $95.1\%$ is managed in most of the cases. This means that the neuro-fuzzy classifier works properly for most of the cases, where both incoherences and coherences take place within the numeric information expressed in 4-tuples.

## 6. Conclusions

This work introduces the problem of content incoherences in document collection, in which connected documents can contain mistakes, wrong or confused cross-contents and the effects of this non coherent documentation are relevant for companies: economic, legal, technical and social damages.

The detection of this type of problem involves extra difficulties with respect to the usual pattern recognition problem: when an incoherence happens in a document, or amongst documents, this depends on the domain documentation and its experts. It is not an objective question, so expert knowledge is needed if success is to be achieved. Here, this expert knowledge is incorporated through a supervised learning procedure supported by a neurofuzzy system.

A global approach is introduced for processing these documents, to detect incoherences: summarization and description of documents is based on heuristics, matching of document contents based on well-known techniques such as the Levenshtein distance or the Cosine similarity, and a supervised learning procedure based on a neurofuzzy system.

Synthetic and real documents summarized by 4-tuples, and matching using the similarity criterion described in the previous section, were used as inputs of the neurofuzzy system for detecting incoherences.

The experiments have shown that the system is able to cope with most cases of coherences and incoherences that can feasibly to take place within a documents set, with a success rate higher than $90\%$. Tests with both synthetically-created cases and real ones have shown that the system is able to learn and detect incoherences by means of the similarities of two 4-tuples holding numeric information.

At present the work is underway concerning the specialization of the FasArt system to be able, not only to detect the existence or not of a numerical incoherence, but also to determine incoherence categories, using the summarization by 4-tuples.

## 7. Acknowledgements

This work has been supported in part by the National Research Agency of Spain (CICYT) through the project DPI2006-15716-C02-02, and the Spanish MITyC through the project TSI-020302-2008-73.

## References

[1] S. Chapman. Sam's String Metrics page, 2006. Available at http://www.dcs.shef.ac.uk/ sam/stringmetrics.html (Accessed Jan.09).

[2] W. W. Cohen, P. Ravikumar, and S. E. Fienberg. A comparison of string metrics for matching names and records. In *Proceedings of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, pages 13–18, Washington DC, USA, 2003.

[3] P. W. Foltz. *Handbook of Latent Semantic Analysis*, chapter Discourse Coherence and LSA, pages 167–184. LEA, 2007.

[4] E. Garcia. Cosine Similarity and Term Weight Tutorial. Mi Islita, Oct 2006. Available at http://www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html (Accessed Jan.09).

[5] N. Koudas, A. Marathe, and D. Srivastava. SPIDER: flexible matching in databases. In *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 876–878, New York, NY, USA, 2005. ACM.

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

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

[8] S. Martín, G. Sainz, and Y. Dimitriadis. Detection of incoherences in a technical and normative document corpus. In *Tenth Int. Conf. on Enterprise Information Systems*, volume Artficial Intelligence and Decission Support Systems, pages 282–287, Barcelona, Spain, June 12–16 2008.

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

[10] G. Sainz Palmero, Y. Dimitriadis, J. Cano Izquierdo, E. Gómez Sánchez, and E. Parrado Hernández. ART based model set for pattern recognition: FasArt family. In H. Bunke and A. Kandel, editors, *Neuro-fuzzy pattern recognition*, pages 147–177. World Scientific Pub. Co., 2000.

[11] G. I. Sainz Palmero and Y. A. Dimitriadis. Structured document labeling and rule extraction using a new recurrent fuzzy-neural system. In *Fifth Int. Conf. on Document Analysis and Recognition, ICDAR' 99*, page 3181, 1999.

[12] G. I. Sainz Palmero, Y. A. Dimitriadis, R. Sanz Guadarrama, and J. M. Cano Izquierdo. Neuro-fuzzy ART based document management system: Application to mail distribution and digital libraries. *Engineering Applications of Artificial Intelligence*, 15(1):17–29, 2002.