HYBRID APPROACH FOR INCOHERENCE DETECTION BASED ON NEURO-FUZZY SYSTEMS AND EXPERT KNOWLEDGE

Susana Martin-Toral[1], Gregorio I. Sainz-Palmero[1,2]
[1]*CARTIF Centro Tecnológico, Parque Tecnológico de Boecillo, parcela 205. 47151 Boecillo, 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: Incoherence, Document corpus, N-tuple, Information retrieval, Neuro-fuzzy system, Expert knowledge, Decision making system.

Abstract: 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. To tackle this situation, this paper proposes the development of an intelligent virtual domain expert, based on summarization, matching and neuro-fuzzy systems, able to detect incoherences about concepts, values, or references, in technical documentation. In this scope, an incoherence is seen as the lack of consistency between related documents. Each document is summarized in the form of 4-tuples terms, describing relevant ideas or concepts that must be free of incoherences. These representations are then matched using several well-known algorithms. The final decision about the real existence of an incoherence, and its relevancy, is obtained by training a neuro-fuzzy system with expert knowledge, based on the previous knowledge of the activity area and domain experts. The final system offers a semi-automatic solution for incoherence detection and decision support.

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 (Martín et al., 2008).

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 (Mingshan and Ching-to, 2002; Afantenos et al., 2005), software companies (Arango, 2003), the legal and law sector (Ruiz, 2002), civil engineering (Martín et al., 2008), 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

Martin-Toral S., I. Sainz-Palmero G. and Dimitriadis Y. (2010).
HYBRID APPROACH FOR INCOHERENCE DETECTION BASED ON NEURO-FUZZY SYSTEMS AND EXPERT KNOWLEDGE.
In *Proceedings of the 12th International Conference on Enterprise Information Systems - Artificial Intelligence and Decision Support Systems*, pages 408-413
DOI: 10.5220/0002966804080413
Copyright © SciTePress

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. An early study of these ideas have been applied in previous works of the authors (Martin et al., 2009). In this paper a new stage in the research and experimentation is shown, presenting advanced results and conclusions.

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. 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 DEFINITION

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* (Martín et al., 2008).

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 (Martin et al., 2009):

- *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 allowed.
- *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 data mining, pattern recognition, semantic analysis, etc. In general, most of them, mainly for extraction and retrieval techniques, are based on the use of heuristic solutions, with similar criteria as in (Krulwich and Burkey, 1997).

3 AN APPROACH FOR INCOHERENCE DETECTION

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 Fig. 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 (Cohen et al., 2003) or the Cosine similarity (Chapman, 2006), 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 (Cano Izquierdo et al., 2001) is used. Although other solutions could be used, this type of systems have been used in previous works for pattern recognition and knowledge extraction (Sainz Palmero and Dimitriadis, 1999; Sainz Palmero et al., 2000; Sainz et al., 2004) 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.

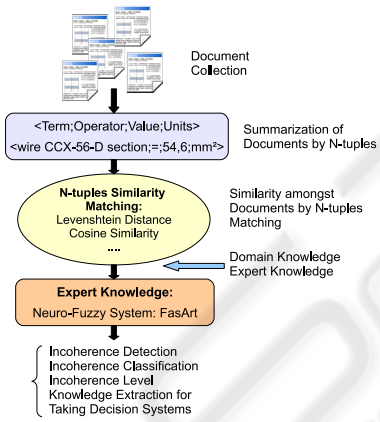

Figure 1: Approach based on neuro-fuzzy system for detection of document incoherences.

3.1 Summarizing Documents by 4-tuples

In this work, information extraction techniques (Berry, 2004) are used to obtain representations that summarize each document of the corpus. Here, the information extraction is based on heuristics (Krulwich and Burkey, 1997), 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 $N = 4$:

$$< \textit{Term ; Operator ; Value ; Units} >$$

$$< \textit{Term ; Operator ; Attribute ; } >$$

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) (Mannila et al., 1997).

An example of real 4-tuples are:

$< \textit{wire CCX-56-D section ; = ; 54,6 ; mm}^2 >$

$< \textit{cover of wire CCX-56-D colour ; = ; green ; } >$

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. 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 incoherences and its relevance. Similarity measures are used to technically define every situation.

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 (Cohen et al., 2003; Chapman, 2006) and Needleman distance (Chapman, 2006).

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* (Garcia, ; Chapman, 2006), *Jaccard similarity* (Cohen et al., 2003; Chapman, 2006), *Dice coefficient* and *Overlap coefficient* (Chapman, 2006).

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 (Koudas et al., 2005) and (Cohen et al., 2003), 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.

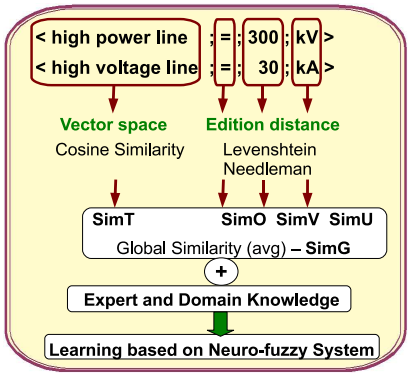

Figure 2: Similarity matching procedure.

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 (see Fig. 2).

3.3 Neuro Fuzzy System FasArt

The FasArt model (Cano Izquierdo et al., 2001; Sainz Palmero et al., 2000) 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 (Cano Izquierdo et al., 2001; Sainz Palmero et al., 2000).

The FasArt system has been used in several previous works (Sainz Palmero and Dimitriadis, 1999; Sainz et al., 2004) 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 neuro-fuzzy system by a set of fuzzy rules, that can be used to generate the virtual expert for automatic incoherence detection.

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. A representative sample of documents from a company of the electric domain containing normative, protocols and operating manuals about usual tasks to be carried out by the company and its partners. This collection consists of 11 documents: 1 main project document referring to the 10 most used normative documents. All these documents were summarized by N-tuples with the most relevant terms or concepts.A total of 3.265 tuples are within this group.
2. Set of synthetic documents containing 5 documents generated by a manual procedure with 29 ideal tuples and several versions of them with different levels of incoherency. A total of 1.185 tuples compose this group.

Once experimental data is ready, matching techniques have been applied element by element to calculate similarity measures (see Fig. 2). Each case has been evaluated by a domain expert, who decides whether if there is incoherence or not, labeling every pattern in a supervised way. The result of the matching stage generates a total of 25 files containing similarity measures for synthetic documents organized in 5 groups (every group with more than 100.000 measures as the input of the system), and 242 files containing similarity measures for real documents organized in 11 groups (every group with more than 1 million measures).

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. The FasArt system has been tuned with respect to vigilance factor ρ and fuzzification rate γ.

On the other hand, the system has been trained and tested by cross validation, using one synthetic or real group for training and the rest synthetic or real groups for testing, and calculating the mean quadratic error of the total trials. Different experiments has been done: 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 real and testing with synthetic documents.

Each experimentation alternative has been evalu-

ated by analyzing the detection error and complexity of the system, through the number of fuzzy rules from the neuro-fuzzy 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 EXPERIMENTAL RESULTS

In Table 1 and Table 2 classification results are shown. Attending to the error rate and the system complexity, two alternatives are possible: 1) Interesting results of classification are obtained when the FasArt system is trained using synthetic tuples and it is tested using real tuples, with an error of 3.58% using Levenshtein distance and 4.44% using Needleman distance. This seems to be coherent, as synthetic tuples have a larger coverage than the real ones, so the synthetic set represents an ideal and theoretical model in the experimentation. 2) Training and testing the system with real tuples offers also good results. The error is by 4.37% using Levenshtein distance, and 2.55% using Needleman distance, and the complexity of the system is smaller than in the previous case. In this situation the system works properly with the documentation of this specific domain and context. But this solution is not general enough to be applied with other documentation and in other context.

Comparing both cases, the first solution seems to be more general, and could work better in more general cases than in the second one, and with a complexity slightly bigger than in the first case. This complexity indicates the number of fuzzy rules we need to collect the expert knowledge, so it is necessary to equilibrate the complexity of the system and the error to obtain a proper solution. This expert knowledge will be reused in order to obtain the virtual expert for incoherence detection. But as the number of rules is high in all the cases, it should be simplify for an adequate use in the expert system.

Table 1: Mean values for the best results for incoherence detection using a FasART classifier.

Similarity	Train	Test	N# Rules	Error
Lev.	Syn.	Syn.	883	6.07%
Lev.	Syn.	Real	799	3.58%
Lev.	Real	Syn.	767	18.12%
Lev.	Real	Real	767	4.37%
Need.	Syn.	Syn.	904	5.92%
Need.	Syn.	Real	800	4.44%
Need.	Real	Syn.	735	78.61%
Need.	Real	Real	735	2.55%

Table 2: Mean values for the worst results for incoherence detection using a FasART classifier.

Similarity	Train	Test	N# Rules	Error
Lev.	Syn.	Syn.	780	15.78%
Lev.	Syn.	Real	914	15.89%
Lev.	Real	Syn.	743	70.90%
Lev.	Real	Real	800	34.97%
Need.	Syn.	Syn.	786	13.45%
Need.	Syn.	Real	873	23.40%
Need.	Real	Syn.	721	85.89%
Need.	Real	Real	721	19.97%

Checking Table 2, the minimum error rate for the worst cases of classification is obtained when the systems are trained and tested using synthetic tuples, a coherent result considering this experimentation as a synthetic scenario.

On the contrary, the maximum error rate of classification is obtained, in all the cases, when real tuples are used to train the system and synthetic tuples are used to test it, with an error near 85% in the worst case. This result is feasible, as only one group of real tuples for training in every trial do not cover the same cases as the synthetic ones.

In the rest of the cases an error rate around 5.5% is obtained for the best cases of classification, validating the experimental stage with a success rate of 95% in most of the situations. 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 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 neuro-fuzzy system in an automatic way.

A global approach is introduced for processing these documents, to detect incoherences: summarization and description of documents is based on heuris-

tics, 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 neuro-fuzzy 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 neuro-fuzzy 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 take place within a documents set, with a success rate higher than 94% in most of the cases. 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 numerical 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 an incoherence, but also to determine incoherence categories, using the summarization by 4-tuples. On the other hand, using this fuzzy approach, it is possible to extract the learnt and subjective expert knowledge from the neuro-fuzzy system, through a set of fuzzy rules that can support a decision making system about this complex and non objective problem.

ACKNOWLEDGEMENTS

This work has been supported in part by the Spanish Industry, Tourism and Commerce Ministry through the project TSI-020302-2008-73.

REFERENCES

Afantenos, S. D., Karkaletsis, V., and Stamatopoulos, P. (2005). Summarization from medical documents: a survey. *Artificial Intelligence in Medicine*, 33(2):157–177.

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

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

Cano Izquierdo, J. M., Dimitriadis, Y. A., Gómez Sánchez, E., and Coronado López, J. (2001). Learnning from noisy information in FasArt and fasback neuro-fuzzy systems. *Neural Networks*, 14(4-5):407–425.

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

Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003). 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.

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

Koudas, N., Marathe, A., and Srivastava, D. (2005). 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. ACM.

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.

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

Martin, S., Arribas, V., and Sainz, G. (2009). Detection of incoherences in a document corpus based on the application of a neuro-fuzzy system. In *Tenth Int. Conf. on Document Analysis and Recognition*.

Martín, S., Sainz, G., and Dimitriadis, Y. (2008). Detection of incoherences in a technical and normative document corpus. In *Tenth ICEIS'08*, volume Artficial Intelligence and Decission Support Systems, pages 282–287, Barcelona, Spain.

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.

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

Sainz, G. I., Fuente, M. J., and Vega, P. (2004). Recurrent neuro-fuzzy modelling of a wastewater treatment plant. *European Journal of Control*, 10:83–95.

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

Sainz Palmero, G. I. and Dimitriadis, Y. A. (1999). 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.