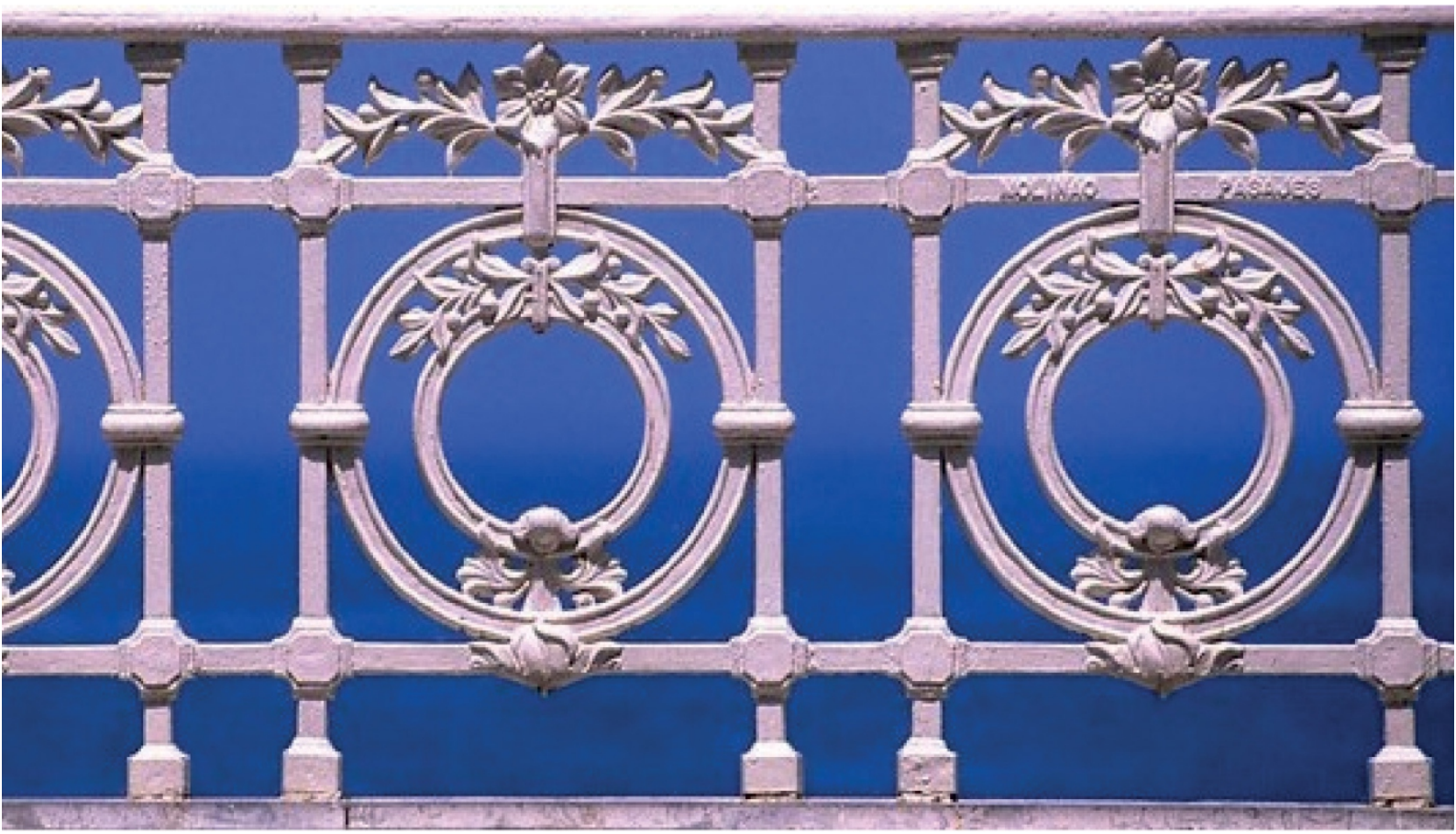
ESTYLF 2016

XVIII Congreso Español Sobre Tecnologías y Lógica Fuzzy



Libro de Resúmenes

XVIII Congreso Español sobre Tecnologías y Lógica Fuzzy

Libro de Resúmenes

Donostia-San Sebastián, 25-27 de mayo de 2016

ORGANIZADORES:

Departamento de Matemática Aplicada.
Universidad del País Vasco/Euskal Herriko Unibertsitatea

EDITORES:

Cristina Alcalde
Humberto Bustince
Francisco Javier Fernández

ENTIDADES PATROCINADORAS:

Universidad del País Vasco/Euskal Herriko Unibertsitatea
Escuela de Ingeniería de Gipuzkoa
Grupo de Investigación de la UPV/EHU «*Sistemas Inteligentes y Energía (SI+E)*»
European Society for Fuzzy Logic and Technology
Red temática de excelencia «*Lógica Difusa y Soft Computing (LODISCO)*»

Red temática de Excelencia:

Lógica Difusa y Soft Computing (LODISCO)

ISBN: 978-84-608-7381-5

Minería de datos, textos y web

Organizadores: M. Dolores Ruiz, Juan Gómez Romero, María José Martín Bautista, Daniel Sánchez

Modelos de optimización y decisión 5D

Organizadores: Dagoberto Castellanos, Carlos Cruz Corona

Problemas de big data con técnicas de soft computing

Organizadores: José Antonio Sanz, Mikel Galar

Open Linked Data Mining for Environmental Scanning*

Juan Gómez-Romero[1], **Fernando Bobillo**[2], **M. Dolores Ruiz**[1], **Miguel Delgado**[1]

[1] *Dep. of Computer Science and Artificial Intelligence, CITIC-UGR, Univ. of Granada*
{jgomez, mdruiz, mdelgado}@decsai.ugr.es
[2] *Aragon Institute of Engineering Research (I3A), Univ. of Zaragoza*
fbobillo@unizar.es

Governmental institutions are increasing their efforts to offer bigger amounts of data as *open data*. By definition, open data can not only be freely accessed in electronic format in the Web, but also reused and redistributed for any purpose. Publishing open data has a two-fold impact to the society. First, it promotes transparency, when referred to public bodies' activities, which is essential in a democratic society. Second, it has social and commercial value, since it can be the enabler of further studies (e.g. independent academics can derive new knowledge) and commercial applications (e.g. socio-economic data can be used in private market analysis). Traditionally, publicly funded statistical offices, such as the Spanish Statistical Office and EUROSTAT, have been the most important open data providers. Nowadays, the definition of open data is widening, and arguably, other sources like social networks and open-licensed digital newspapers can be as well included.

Environmental scanning is defined as the process of acquisition and use of information about events, trends and relationships in an organization's external environment to support decision-making and planning [1]. The wide availability of open data has renewed the interest in this topic, in which Data Mining plays a key role regarding automatic discovery of non-trivial, new and valid knowledge from data. We identify here three challenges of open data mining for environmental scanning:

Data heterogeneity. Using different information sources is essential to provide a wider context to understand the whole scenario. Data combination implies incompatible formats, which must be expressed in a common language, and more importantly, semantic heterogeneity, which requires the alignment of diverse conceptual schemas based on information meaning rather than data format. This is particularly difficult if unstructured sources are to be considered; e.g. texts in natural language.

Less structured process. The typical data mining process encompasses four steps: data collection and cleaning (pre-processing), filtering (selection), exploration and model building (analysis), and visualization (for data description and prediction). To some extent, open data mining invert this process, because it implies the collection of data before determining an explicit purpose. Therefore, we may be more interested in exploratory analysis than model building, which would not require a thorough pre-processing. Conversely, it may happen that the pre-processing stage requires more effort before starting a detailed analysis aimed at decision-making.

Imprecision and vagueness. Open data makes it necessary to reinforce the capabilities of mining algorithms to discern between relevant and irrelevant information, and to be

* This work is partly funded by the Univ. of Granada (Programa de Proyectos para la incorporación de jóvenes doctores a nuevas líneas de investigación), the Spanish Ministry of Economy and Competitiveness (CICYT project TIN2013-46238-C4-4-R), and the Andalusian Government (P11-TIC-7460).

resilient to noise, uncertainty, and outliers. In this regard, Fuzzy Logic methods have proved to be more robust than conventional approaches [2]. In non-fuzzy methods, small changes in the inputs result in major changes in the outputs, whereas in fuzzy methods, changes are more gradual.

Semantic Web technologies, based on ontological and graph-based knowledge representations, provide an infrastructure for publishing, storing, retrieving, reusing, integrating, and analyzing open and linked data [3]. The standard Semantic Web Languages (RDF, RDFS, OWL) allow us to express data and metadata in the form of triples (subject, property, object) with formal properties, and more importantly, to easily link pieces of information. The Linked Data initiative purposely focus on connecting related information on the Web [4]. For example, DBPedia (a semi-structured version of Wikipedia) is linked to GeoNames (a geographical knowledge base). Specialized vocabularies based on RDF and OWL have been proposed to characterize, find and use open datasets: VoID, to express metadata about datasets; PROV-O, to define data provenance; Data Cube Vocabulary, for statistical data and metadata exchange; and so forth. Queries can be issued to distributed RDF sources, published in triplestore repositories, by using the SPARQL standard query language. All these technologies favor the creation of rich information contexts, with wider scope than the initial target data, at a low cost.

To illustrate the potential of open data mining, we have developed an example based on EUROSTAT, DBPedia and GeoNames open data. From EUROSTAT, we have identified a dataset about unemployment in European regions, encoded with the Data Cube vocabulary and linked to GeoNames and DBPedia, and we have retrieved unemployment rates for "all the European regions where German is an official language in 2000-2012". Let us notice that the regions are defined in GeoNames, the official languages in DBPedia, and the unemployment rates in EUROSTAT. We have applied a fuzzy clustering algorithm to characterize groups of regions with similar evolution, which allows the use of imprecise cluster boundaries. The results suggest that the unemployment trends mostly depend on the initial rates and the changes in the first two years of the economical crisis, regardless of the country. We could incorporate other sources to the process, and refine the conditions to explore specific scenarios with small effort.

We consider that the use of ontological descriptions for linked data is one of the most promising directions for future work. Both the input data and the results –the clusters, in our example– could be semantically described with meaningful ontological expressions. Moreover, we could use a fuzzy ontology, in which concepts and relations can be imprecise. Another area of interest is the utilization of automatic ontology alignment algorithms in the pre-processing stage, in order to locate and link similar datasets.

References

[1] C.W. Choo: *The art of scanning the environment*. Bulletin of the American Society for Information Science and Technology 25(3) (1999), 21-24.

[2] E. Hüllermeier: *Fuzzy methods in machine learning and data mining: Status and prospects*. Fuzzy Sets and Systems 156(3) (2005), 387-406.

[3] K. Janowicz, F. van Harmelen, J. Hendler, P. Hitzler: *Why the data train needs semantic rails*. AI Magazine 36 (1) (2014).

[4] C. Bizer, T. Heath, T. Berners-Lee: *Linked Data - The story so far*. International Journal on Semantic Web and Information Systems 5(3) (2009), 1-22.