Introduction to the Special Issue on Mining the Humanities: Technologies and Applications ☆

In the era of the current digital revolution our society is based on the so-called digital industry surrounded by a high-tech global economy. Data and information play a key role building relationships with a range of organizations working on data; the latter taking a variety of forms, such as big, small, linked and open. This information industry allows humans to explore their personalized needs, therefore simplifying the procedure of interacting with the digital world. As it seems that almost any conversation on information and communication these days inevitably includes the term "data" at some point, it is more than evident that their interpretation must be contextual and aligned with everyday's human needs and actions. As a consequence, a huge amount of heterogeneous human-related data is generated, enriched and exploited for creating new research opportunities and challenges.

Still, the nature of humanistic data can be multimodal, semantically heterogeneous, dynamic, time and space-dependent, and highly complicated. Translating humanistic information, such as data deriving from human behavior analysis, state of mind interpretation, artistic creation, linguistic utterance or even learning and genomic information into numerical or categorical low-level data, is a significant challenge on its own. Computer science researchers do work on new techniques that would be appropriate to deal with this type of data. Some interesting novel propositions are included in the following, whereas interesting existing ones are being adapted to match the special characteristics of the new domain. Last, but not least, the close link with artificial intelligence methodologies and applications is rather self-evident: research in this area is important to combine the fields of semantics and knowledge engineering within artificial intelligence framework, mainly because of the overwhelming amount and different types of disseminated information.

In this Special Issue, we present original papers that tackle challenges or issues relating to bringing together interdisciplinary approaches that focus on the application of innovative, as well as existing data matching, fusion and mining and knowledge discovery and management techniques to data derived from a variety of humanistic sciences ranging from linguistic to historical and educational ones. A total of twenty one (21) papers have been selected for this Special Issue and report the latest advances on the technologies, algorithms, models, standards and applications in the aforementioned topics.

1. The Articles

Such was the wide public acceptance of this initiative that the Call for Papers of the Special Issue received an extraordinary strong response from various communities, including multimedia semantics, user modeling, personalization, artificial intelligence and knowledge analysis. Sixty seven (67) high quality contributions addressing related theoretical and practical aspects of the call were initially summoned and a total of twenty one (21) papers were finally accepted following a rigorous two-stage reviewing round, coordinated by the guest editors. In the following, after providing a topic-related clustering of accepted papers into four (4) main groups according to their main field of interest, a short description for each paper is provided, summarizing the aims of each article and how the work described is related to the Special Issue's topics.

1.1. Humanistic Data Mining and Information Retrieval Algorithms

The first paper of the Special Issue, entitled "*Social Network Regularized Sparse Linear Model for Top-N Recommendation*", prepared by *Xiaodong Feng, Ankit Sharma, Jaideep Srivastava, Sen Wu and Zhiwei Tang*, introduces social recommendation techniques to employ user's social connections. It focuses on developing effective methods to utilize social network information for Top-N recommendation, including the social network regularized Sparse Linear Model (SocSLIM) with its extensions incorporating local learning (LocSocSLIM). Experimental results demonstrate that SocSLIM effectively uses the social information to outperform the state-of-the-art methods by a significant percentage, whereas the utilization of local weight learning extension LocSocSLIM significantly improves the overall efficiency.

Xenophon Evangelopoulos, Victor Giannakouris-Salalidis, Lazaros Iliadis, Christos Makris, Yannis Plegas, Antonia Plerou and Spyros Sioutas in "*Evaluating Information Retrieval using Document Popularity: An implementation on MapReduce*" present an innovative approach to user behavior prediction. They propose a new evaluation metric for information retrieval systems, which employs two relevance factors, namely a relevance judgment coupled together with a popularity grade that represents users' vote for a document. The authors show through an extensive evaluation paradigm that this new metric performs better than other evaluation metrics when expressing user behavior.

Antonela Tommasel, Alejandro Corbellini, Daniela Godoy, and Silvia Schiaffino in their paper "Personality-aware followee recommendation algorithms: An empirical analysis" aim to tackle the fundamental

problem of deciding who to follow in micro-blogging sites. Their study assesses the impact of personality in the accurate prediction of followees, beyond simple topological and content-based factors. It analyses whether user personality could condition followee selection by combining personality traits with the most commonly used followee predictive factors. Experimental results showed that an accurate appreciation of such predictive factors tied to a quantitative analysis of personality is crucial for guiding the search of potential followees, and thus, enhance overall recommendations.

The fourth paper of the issue entitled "*Discovering similar Twitter accounts using semantics*", prepared by *Gerasimos Razis and Ioannis Anagnostopoulos* proposes to examine a popular social network, Twitter, towards the introduction of a novel methodology for discovering and suggesting similar Twitter accounts, based entirely on their disseminated content in terms of the Twitter entities used. The methodology of the authors is based exclusively on semantic representation protocols and related technologies, whereas an ontological schema is also described within the paper towards the semantification of Twitter accounts and their entities.

Basilis Charalampakis, Dimitris Spathis, Elias Kouslis and Katia Kermanidis in their paper entitled "*A comparison between semi-supervised and supervised text mining techniques on detecting irony in Greek political tweets*" describe a classification schema for irony detection in Greek political tweets. Their hypothesis states that humorous political tweets could predict actual election results. The irony detection concept is based on subjective perceptions, so only relying on human-annotator driven labor might not be the best route. The proposed approach relies on limited labeled training data, thus a semi-supervised approach is followed, where collective-learning algorithms take both labeled and unlabeled data into consideration. The hypothesis is evaluated via a correlation study between the irony that a party receives on Twitter, its respective actual election results during the Greek parliamentary elections of May 2012, and the difference between these results and the ones of the preceding elections of 2009.

Aicha Boutorh and Ahmed Guessoum in the paper entitled "*Complex Diseases SNP Selection and Classification by Hybrid Association Rule Mining and Artificial Neural Network-based Evolutionary Algorithms*", provide an overview on Single Nucleotide Polymorphisms (SNP) data classification as the latter have been shown to be implicated in various human diseases. According to their research one of the major problems related to SNP sets is the high number of features and the small number of samples, which makes the classification task rather complex. In this paper they propose a new hybrid intelligent technique based on Association Rule Mining (ARM) and Neural Networks (NN), which uses Evolutionary Algorithms (EA) to deal with the dimensionality problem. On the one hand, ARM optimized by Grammatical Evolution (GE) is used to select the most informative features and to reduce the dimensionality by parallel extraction of associations between SNPs in two separate datasets of case and control samples. On the other hand, and to complement the previous task, a NN is used for efficient classification. The Genetic Algorithm (GA) is used for setting up the parameters of the two combined techniques. The proposed GA-NN-GEARM approach has been applied on four different SNP datasets obtained from the NCBI Gene Expression Omnibus (GEO) website, whereas the created model has reached a high classification accuracy.

Evaggelos Spyrou and Phivos Mylonas in their paper entitled "*A survey on Flickr multimedia research challenges*" focus on the Flickr social network and review the recent advances in the fields of humanistic digital content analysis and management techniques, as well as its meaningful interpretation and exploitation. They demonstrate and discuss some of the major research challenges in the area including current state-of-the-art approaches with respect to the humanistic data collection and interpretation research fields dealing with aspects such as multimedia information retrieval,

(semi-) automatic tag manipulation, travel applications, semantic knowledge extraction, human activity tracking, and related benchmarking efforts. At the end of this survey, the authors also discuss the main challenges that are to be identified and propose a number of future research directions for the interested parties.

The last paper of this group entitled "*On the problem of early detection of users interaction outbreaks via stochastic differential models*" is proposed by Markos Avlonitis and tackles the challenging human-computer interaction problem of early detection of users' interaction outbreaks. More specifically a new stochastic differential model for outbreaks detection of an arbitrary time series is proposed in his paper, whereas the discussed methodology introduces an explicit stochastic differential law for the evolution of outbreaks, i.e., mainly a limiting case of the stochastic generalization of the well-known Verhustl model. The key feature of the proposed algorithm is a non-linear transformation of the corresponding population state variable. At the end, the detection of an outbreak regime in time is connected with the detection of a time regime over which the transformed variable obeys a random walk stochastic evolution. An application of the proposed algorithm is given for the problem of outbreaks in users' video interactions, where the validity and usefulness of the algorithm is demonstrated.

1.2. Pattern Matching Approaches

The Special Issue continues by introducing three papers dealing with pattern matching approaches, namely the first one entitled "*Design Reuse in the Conceptual Schema of CMSs: a pattern-based Evaluation Approach*" by *Vassiliki Gkantouna, Spyros Sioutas, Athanasios Tsakalidis, Giannis Tzimas and Emmanouil Viennas* addresses the inspection and evaluation of design reuse in the conceptual schema of CMS-based Web applications in order to improve their design quality. The authors utilized WebML as the design platform of their proposed approach and they capture design reuse by detecting all the recurrent design structures within the hypertext schema of an application. The proposed methodology is exploited towards the automatic extraction of the hypertext schema of an application which is then submitted to a pattern-based analysis, in order to detect all the incorporated recurrent patterns implying design reuse. Finally, the authors apply a number of metrics on the identified patterns to evaluate the appropriateness of reuse which results in categorizing them as effective or poor design solutions.

The second paper was written by *Maxime Crochemore, Costas S. Iliopoulos, Ritu Kundu, Manal Mohamed and Fatima Vayani* and is entitled "*Linear Algorithm for Conservative Degenerate Pattern Matching*". It studies and discusses the matching problem of conservative degenerate strings and presents an innovative algorithm that for given degenerate strings is able to identify accompanying occurrences in linear time. More specifically the authors used the novel technique of substituting the non-solid symbols in the given degenerate strings with unique solid symbols, which let them make use of the efficient approximate pattern search solution for solid strings to get an efficient solution for degenerate strings.

The third paper of the group entitled "*Associating ω -automata to Path Queries on Webs of Linked Data*" and prepared by *Konstantinos Giannakis, Georgia Theocharopoulou, Christos Papalitsas, Theodore Andronikos and Panayiotis Vlamos* discusses the issue of *the emergence of semantically enriched data sets along with the elegant way they are interconnected through the concept of Linked Data. Within this paper authors show that there are queries that cannot be expressed neither by standard SPARQL nor its expansion with regular path queries. These queries assume that the Web of Linked Data can be infinite and therefore classic queries expressed in terms of SPARQL need to be revisited. The authors propose a novel method associating a variant of automata that read infinite inputs, namely the ω-automata, to each query on this infinite structure.*