Application of an Automatic Data Alignment & Structuring System for Intercultural Consumer Segmentation Analysis

Fumiko Kano Glückstad
Dept.of International Business Communication, Copenhagen Business School, Dalgas Have 15, Frederiksberg, Denmark

Keywords: Ontology Alignment, Knowledge Structuring, Nonparametric Bayesian Relational Modelling, Cross-cultural Data Analysis, Opinion Survey Analysis, Consumer Segmentation.

Abstract: This position paper introduces a conceptual framework of our ambitious international research project where the aim is extraction and alignment of heterogeneous consumer segment structures across a multiplicity of markets and cultures. We argue that an automatic data alignment and structuring system employing a non-parametric Bayesian relational modelling is an ideal approach that can address challenges in the conventional cross-cultural data analysis. The paper presents an example of our preliminary work that applies this approach to the analysis of opinion survey responses given by male populations in Sweden and Japan. The framework successfully extracts groups of males who express similar but also dissimilar response patterns from the two selected countries. Based on these preliminary studies, the paper discusses potential contributions and future challenges of the international consumer analysis project.

1 INTRODUCTION

The establishment of semantic relations across boundaries of e.g. organizational-, linguistic-, or societal communities (Fensel et al., 2004) has been a central topic of the ontology matching discipline (Euzenat and Svaiko 2007) and substantial numbers of ontology matching frameworks have been introduced in recent years (Berlin and Motro, 2002; Cheng et al., 2008; Doan et al., 2004; Euzenat, 1994; Stumme and Mädche, 2001; Wang et al., 2004; Mørup et al., 2014). However, a fully automatic ontology alignment has remained a highly challenging task since it typically involves similarity computations between objects belonging to different knowledge systems. As indicated in previous works (Isaac et al., 2007; Pirrò and Seco, 2008; Pirrò and Euzenat, 2010; Ngo et al., 2013; Cross et al., 2013; Glückstad et al., 2014; Mørup, 2014), similarity measures influence on the performance of ontology alignment tasks. Mørup et al. (2014) addresses this ambiguity by introducing a new framework applying the nonparametric Bayesian relational modeling to statistically analyze feature matches of all combinations of objects belonging to different knowledge systems, i.e., to align them while jointly partitioning them into clusters. This new way of analyzing "relatedness" between objects belonging to different knowledge systems engenders an additional ability to automatically align three or more ontologies simultaneously (Mørup et al., 2014).

Whereas this ability to automatically align and structure multiple datasets contributes substantially to the knowledge engineering and ontology development discipline (e.g. Glückstad et al., 2014; Mørup et al. 2014), such ability could also be useful in other research disciplines which requires the analysis of "relatedness" of objects across multiple datasets. This position paper introduces our ambitious international research project proposal which sheds light on the potential contributions of such automatic data alignment and structuring framework to other research disciplines, namely, the cross-cultural consumer segment analysis.

The next section first introduces the conceptual framework of the proposed project followed by potential contributions of the automatic data alignment and structuring approach to the extraction of heterogeneous consumer segment structures across a multiplicity of cultures. Finally, we briefly show an example of our preliminary study examining the applicability of Mørup's framework to analyse cross-cultural opinion survey responses.

Glückstad, F..
Application of an Automatic Data Alignment & Structuring System for Intercultural Consumer Segmentation Analysis.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 2: KEOD*, pages 251-256
ISBN: 978-989-758-158-8

2 CONCEPTUAL FRAMEWORK

International marketing professionals who have extensive professional experience in the Far East markets are aware that consumers in China or Japan cannot simply be categorized by a standardized consumer segmentation model universally used in the Western marketing practice. They need scientific and empirical documentation comprehending similarities and differences of consumer segments across boundaries of societies. Cleveland et al., (2011a) emphasizes that: "a proportion of individuals worldwide develop bicultural identities: one based in local traditions combined with an identity connected to an emerging global culture" and "as corporations globalize, the key challenge for managers is to institute an effective marketing orientation across a composite of cultures".

Values are crucial for explaining the motivational basis of human attitudes and behaviour (Schwartz, 2012). Individuals act, make decisions, emotionally react to advertisements, purchase and assess products based on their identity-based motivations. However, the identity formation and thereby value formation of modern consumers are becoming increasingly complex due to their belonging to local, national and global communities accessible via e.g. contemporary media technologies. Hence, it is crucial to map out and comprehend the complexity of consumers, i.e. the heterogeneity of consumer segment structures across societies (Cleveland et al., 2011a; Cleveland et al., 2011b). Our hypothesis is that Schwartz's Theory of the Ten Basic Human Values (STBH: Schwartz, 2012) is a useful measurement scale for contrasting consumers across a multiplicity of cultures (Fischer and Schwartz, 2011). We hypothesize that, by analysing the existing data from the World Value Survey (WVS) including the question items of STBH, it is feasible to develop a new measurement scale that can effectively extract heterogeneous consumer segment structures across cultures.

From a technical aspect, our hypothesis is that the nonparametric Bayesian relational modelling is a powerful and suitable approach (Schmidt and Mørup, 2013; Mørup and Schmidt, 2012) for extracting and aligning consumer segments existing across a multiplicity of markets. The model jointly partitions consumers in the respective markets into segments representing either universal value patterns across cultures or culturally specific value patterns. By inspecting behavioural data connected to members of the value-based consumer segments extracted by the proposed approach, the project will eventually be able to infer the preferences of targeted consumers and their predicted behaviours. Accordingly, we eventually intend to develop a data-analytic platform and test its usability by questioning:

- What type of consumer segments will potentially be interested in [specific products/services] in selected foreign markets?
- How can the potential consumer segment of [specific products/services] be characterized? Are the characteristics universal across cultures or culturally specific?
- Are the newly emerged segments, such as the "Millenials", a transnational- or a nationally specific phenomenon across these markets?

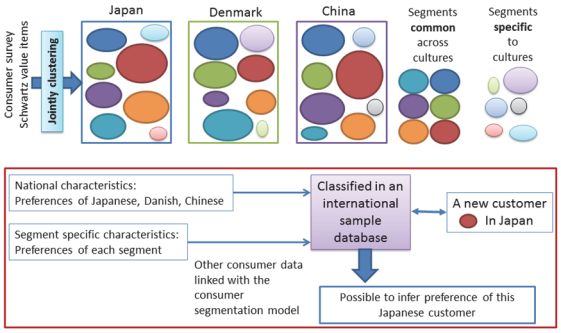


Figure 1: The conceptual framework.

Figure 1 illustrates the conceptual framework of the intercultural consumer segmentation system. In this framework, feature vectors corresponding to responses to the question items of STBH given by each consumer are used for segmenting consumers. By analysing the "relatedness" between responses of consumers (feature vectors) within and across cultures, a methodological framework employing the nonparametric Bayesian relational model simultaneously extract consumer segments similar but also dissimilar across cultures. The next section briefly reviews this joint clustering mechanism.

3 METHODOLOGICAL FRAMEWORK

The automatic data alignment framework introduced in Mørup et al., (2014) employs a statistical model belonging to a family of stochastic block-models widely used in the social network analysis (Faust and Wasserman, 1992; Wasserman and Anderson, 1987) and also applied in the complex brain connectivity network (brainconnectivity.compute.

dtu.dk, Mørup et al., 2010; Mørup and Schmidt, 2012; Schmidt and Mørup, 2013; Schmidt et al., 2012). The principle of the stochastic block-model is to partition a set of objects into clusters whose members are stochastically equivalent and possess similar relations to other clusters to be extracted in the partitioning process. In Mørup's framework (Mørup et al., 2014), an extended version of the so-called Infinite Relational Model (IRM: Kemp et al., 2006; Xu et al., 2006), a family of binary stochastic block-models is employed. As is explained in Kemp et al. (2006), the uniqueness of the IRM is to automatically extract an appropriate number of clusters based on the so-called Chinese Restaurant Process (CRP: Aldous, 1985; Pitman, 2002).

Whereas the original IRM is suitable solely to the ontology learning, the extended IRM, called the multinominal IRM (mIRM) in Mørup et al. (2014), is based on a multinominal observation likelihood and is suitable to align multiple datasets. A distinctive characteristic of the mIRM framework is to first count all combinations of binary feature matches between objects, which will be used for the alignment across datasets during the partitioning process employed in the mIRM. Specifically, when aligning two datasets whose objects consist of respective feature vectors, four combinations of matches are counted: i) 1-1 match means that an object from System A and an object from System B possess the same binary feature; ii) 1-0 match means that an object from System A possess a feature that is not possessed by an object from System B; iii) 0-1 match means that an object from System B possess a feature that is not possessed by an object from System A; and iv) 0-0 match means that a feature is possessed neither by an object from System A nor by an object form System B. Following the same principle, the alignment of three systems counts eight combinations of matches: i) 1-1-1; ii) 1-1-0; iii) 1-0-1; iv) 1-0-0; v) 0-1-1; vi) 0-0-1; vii) 0-1-0; and viii) 0-0-0. This further implies that it is feasible to align four or more systems by counting all possible combinations of matches, i.e. 16 combinations in case of four systems, 32 combinations for five systems and so forth. This count statistics of binary feature matches replaces similarity measures discussed in the previous literatures (Isaac et al., 2007; Pirrò and Seco, 2008; Pirrò and Euzenat, 2010; Ngo et al., 2013; Cross et al., 2013; Glückstad et al., 2014).

The count statistics obtained by this process is further used for simultaneously partitioning and aligning sets of objects between multiple systems. For example, in the alignment of two systems, the mIRM is defined as the following generative process where objects in the two systems are simultaneously partitioned into z and w clusters according to the CRP.

$$\boldsymbol{z} \sim CRP(\alpha^{(1)}), \boldsymbol{w} \sim CRP(\alpha^{(2)}),$$
$$\boldsymbol{\eta}_{lm} \sim Dirichlet(\boldsymbol{\eta}_0),$$
$$\boldsymbol{f}(\boldsymbol{u}_i, \boldsymbol{v}_j) \sim Multinomial(\boldsymbol{\eta}_{z_i w_j}, N_{ij}).$$

The probabilities that the four types of matches are observed between an extracted cluster l in System A and cluster m in System B are defined as $\boldsymbol{\eta}_{lm}$ according to the Dirichlet distribution. Finally, the count statistics of the four types of matches $\boldsymbol{f}(\boldsymbol{u}_i, \boldsymbol{v}_j)$ between an object i in System A and an object j in System B are drawn from the multinominal distribution where the between-cluster probabilities for each of the four combinations of matches are defined as $\boldsymbol{\eta}_{z_i w_j}$. For those interested in further details of the algorithms, please refer to: Kemp et al., (2006; 2010) for the original IRM; Schmidt and Mørup (2013) for the general review of the nonparametric relational modelling; and Mørup et al., (2014) for the mIRM.

Some of the major contributions demonstrated by Mørup et al., (2014) are that i) the automatic alignment of two as well as three knowledge systems has been achieved and ii) the alignment of systems have not been influenced by the choice of similarity measure. While Mørup et al. demonstrated the alignment of knowledge systems across two and three legal systems (educational systems of two and three countries defined by UNESCO), the present work applies this analytical framework to a new type of dataset, namely cross-cultural opinion survey responses. The latter analysis demonstrates how the alignment results of the four types of matches are interpreted in the domain of opinion survey analysis.

4 PRELIMINARY STUDY

4.1 Datasets

For applying the mIRM framework to a cross-cultural opinion survey analysis, the 6th Wave of World Value Survey (WVS) datasets downloadable from the WVS organization web-site are employed. The data used in this analysis are collected from 557 Swedish and 954 Japanese males who responded to all ten question items V79-89 corresponding to the Portrait Values Questionnaire (PVQ) developed by Schwartz (2012). For example, the portrait of "Self-Direction value (V79)" describes: "*Thinking up new*

ideas and being creative is important to him. He likes to do things in his own original way." Accordingly, respondents must select an answer from the six-level Likert scale: 1. *Very much like me*; 2. *Like me*; 3. *Somewhat like me*; 4. *Little like me*; 5. *Not like me*; and 6. *Not at all like me*. In this analysis, these six answer categories are semantically divided into a positive answer covering 1-4 and a negative answer including 5 and 6. This means that there are 2^{10} patterns for answering these ten PV Q questions.

4.2 Analysis of the Aligned Results

The mIRM has been run 20 times with 5000 iterations in this analysis. The Normalized Mutual Information (NMI) scores for the 20 runs are 0.91 for the Japanese and 0.92 for the Swedish. Figure 2 displays that the mIRM extracted in total 81 and 58 clusters respectively from the Japanese and Swedish datasets.

Figure 2 displays four plots named as 1-1 matches, 1-0 matches, Jaccard and SMC (Simple Matching Coefficient). The two plots names as "1-1 matches" and "1-0 matches" are examples of the clusters extracted by the mIRM estimating an appropriate number of clusters to be extracted based on the count statistics of the four combinations of the matches (1-1, 1-0, 0-1, and 0-0). The intersections uniformly highlighted with yellow in the plot of "1-1 matches" indicate that all cluster members between the Japanese (Y axis) and Swedes (X axis) agree either positive or negative on all the ten PVQ question items.

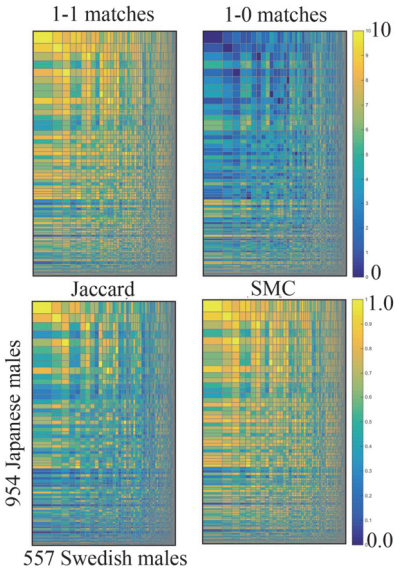
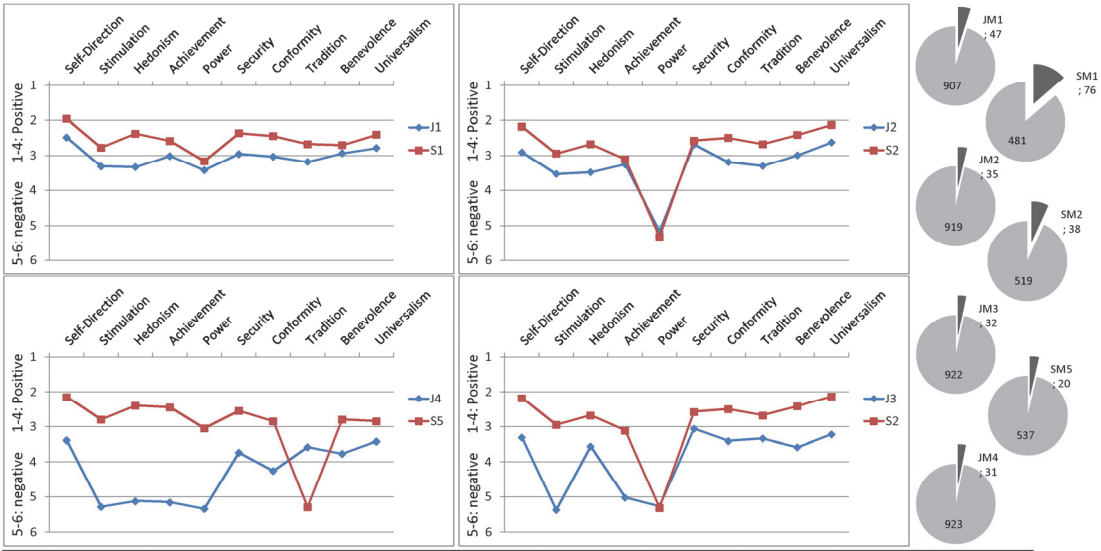


Figure 2: Clusters extracted and aligned between Japanese and Swedish males.



Cluster ID	Self-Direction		Stimulation		Hedonism		Achievement		Power		Security		Conformity		Tradition		Benevolence		Universalism	
	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.
JM1	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%
SM1	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%
JM2	100%	0%	100%	0%	100%	0%	100%	0%	0%	100%	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%
SM2	100%	0%	100%	0%	100%	0%	100%	0%	0%	100%	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%
JM4	90%	10%	0%	100%	0%	100%	0%	100%	0%	100%	81%	19%	52%	48%	100%	0%	81%	19%	97%	3%
SM5	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%	0%	100%	100%	0%	100%	0%
JM3	100%	0%	0%	100%	100%	0%	0%	100%	0%	100%	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%
SM2	100%	0%	100%	0%	100%	0%	100%	0%	0%	100%	100%	0%	100%	0%	100%	0%	100%	0%	100%	0%

Figure 3: Cluster mean scores of the ten question items, size of the clusters, and the proportion of the positive and negative answers indicated by the members of the respective clusters.

On the other hand, an intersection highlighted with colours closer to yellow in the plot of "1-0 matches" indicate that a Japanese and a Swede belonging to the intersection disagree on many of the ten question items. The plots named as "Jaccard" and "SMC" displays the levels of similarity scores computed between Japanese and Swedish respondents belonging to the respective clusters extracted by the mIRM. A noteworthy point is that, if these similarity measures are used for co-clustering the same respondents, the clustering results will be influenced by the distribution of scores between the highest "1.0" and the lowest "0.0" as shown in (Glückstad et al., 2014).

Figure 3 further inspects the actual responses of the cluster members to the ten PVQ question items (V79-89). The four plots in Figure 3 show the overall results for the STBH values for specific cluster combinations: J1-S1; J2-S2; J4-S5; and J3-S2. For example, Figure 3 indicates that all members belonging to the Japanese cluster 1 (JM1: 47 respondents) and the Swedish cluster 1 (SM1: 76 respondents) answered positively to all ten PVQ questions. On the other hand, all members in JM2 and SM2 responded negatively to the question regarding the "Power" value. These two combinations of clusters are the ones connected via the uniformly yellow intersections in "1-1 matches" of Figure 2. In the case of JM3-SM2 combination, members within the respective clusters all agree on either negative or positive responses. However, disagreement appears between JM3 and SM2 in "Stimulation" and "Achievement" values. These clusters are linked with the intersection uniformly coloured with Orange in "1-1 matches" of Figure 2. Finally, in the combination between JM4 and SM5, members of JM4 disagree on several questions such as "Conformity" value within JM4 as well as disagree with SM5 on PVQ items such as "Stimulation" value. These clusters are linked with the green coloured intersection, which indicate that the relation between the two clusters is rather ambiguous, i.e. neither similar nor dissimilar.

5 CONCLUDING REMARKS

The previous section demonstrated the joint clustering of two datasets compiled from the ten PVQ question items representing STBH values. The mIRM framework applied in this work has extracted groups of people who express similar but also dissimilar response patterns from the respective datasets. The extracted clusters are aligned between the two culturally specific datasets, of which between-cluster relations are observable in the matrices displayed in Figure 2. A noteworthy point is that these matrices illustrate the overall similarity relations of respondents within- and across-clusters based on the differentiated colours. This means that these similarity relations might be further used for effectively visualizing the hierarchical structures of the extracted clusters by use of existing ontology learning technologies.

Another important remark is that the application of the ontology engineering to the conventional cross-cultural data analysis may add substantial values to this specific research domain. Traditionally, cross-cultural data analyses in Social Sciences have rather focused on the analysis of mean values of predefined segments such as "a group of Japanese female teenagers" and based on researcher's hypotheses and comparisons across cultures. By use of ontology engineering technologies that can focus on structuring relations between individuals based on their feature vectors, the heterogeneous segment structures can be extracted in a data-driven manner. Such an approach is particularly needed for comprehending and comparing heterogeneity across societies. Recalling that in our contemporary society, individuals' values are shaped and influenced by a multiplicity of layers ranging from the mono-cultural (local), multicultural (regional) and transcultural (universal) layers in the modern global world. The present work has demonstrated a pre-cursor for challenging such new and exciting research endeavours.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Profs. Morten Mørup and Mikkel N. Schmidt who have originally developed the mIRM model and the related analytical toolbox used in the present work. I also thank Microsoft Denmark and Techila Technologies who provided a high performance cloud-computing infrastructure that has successfully reduced the computational time in the present work from 180 hours (under the "non" parallel computing environment using CPU: Intel Core i5M480 @2.67GHz, Memory: 4 GB) to just 40 minutes.

REFERENCES

Aldous, D. 1985. Exchangeability and Related Topics. in

E´Cole D'e´Te´ De Probabilite´S Desaint-Flour XIII1983, 1–198

Berlin, J. & Motro, a. 2002. Database Schema Matching using Machine Learning with Feature Selection. in Proceedings of the 14th International Conference on *Advanced Information Systems Engineering, Vol. 2348 of Lecture Notes in Computer Science*, 452–466

Cheng, C.P., Lau, G.T., Law, K.H., Pan, J. & Jones, a. 2008. Regulation Retrieval using Industry Specific Taxonomies. in *Artif Intell Law 16, Springer*. 277–303.

Cleveland, M., Erdoğan, S., Arıkan, G., Poyraz, T. 2011a. Cosmopolitanism, Individual-Level Values and Cultural-Level Values: A Cross-Cultural Study. in *Journal of Business Research 64*, 934–943.

Cleveland, M., Papadopoulos, N., Laroche, M. 2011b. Identity, Demographics, and Consumer Behaviors: International Market Segmentation across Product Categories, in *Intl. Marketing Review, 28(3)*, 244-266

Cross, V., Yu, X. and Hu, X. 2013. Unifying Ontological Similarity Measures: a Theoretical and Empirical Investigation. in *International Journal of Approximate Reasoning, Vol.54*, 861–875.

Doan, A.H., Madhavan, J., Domingos, P. & Halevy, a. 2004. Ontology Matching: a Machine Learning Approach. in: *Staab S, Studer R (Eds) Handbook on Ontologies, Chapter 18. Springer, Berlin,* 385–404

Euzenat, J. 1994. Brief Overview of T-Tree: the Tropes Taxonomy Building Tool. in *Proceeding of the 4th ASIS SIG/CR Workshop on Classification Research*, 69–87

Euzenat, J. & Shvaiko, P. 2007. *Ontology Matching. Springer, Berlin*.

Faust, K. & Wasserman, S. 1992. Blockmodels: Interpretation and Evaluation. in *Social Networks, 14 (1-2)*, 5-61

Fensel D., Davies, J., Bussler, C. and Studer R. (Eds.) 2004. the Semantic Web: Research and Applications. in *First European Semantic Web Symposium, Heraklion, Crete, Greece, Springer-Verlag, Berlin*.

Fischer, R. Schwartz, S.H. 2011. Whence Differences in Value Priorities? Individual, Cultural, or Artifactual Sources. in *Journal of Cross-Cultural Psychology, Vol 42*, 1127-1144

Glückstad, F.K., Herlau, T., Schmidt, N. M., & Mørup, M. 2014. Cross-Categorization of Legal Concepts across Boundaries of Legal Systems: in Consideration of Inferential Links. in *Artif Intell Law 22 (1), Springer,* 61-108.

Isaac, A., Meij, L.V.D., Schlobach, S. & Wang, S. 2007. an Empirical Study of Instance-based System Matching. in *the Semantic Web, Lecture Notes in Computer Science, Vol.4825,* 253–266.

Kemp, C., Tenenbaum, J.B., Griffiths, T.L. Yamada, T. & Ueda, N. 2006. Learning Systems of Concepts with an Infinite Relational Model. in *Proc. the 21st National AAAI Conference on* 1:381-388.

Mørup, M., Glückstad, F.K., Herlau, T. & Schmidt, N. M. 2014. Non Parametric Statistical Structuring of Knowledge Systems using Binary Feature Matches. in *Proceedings of 2014 IEEE International Workshop on Machine Learning for Signal Processing*

Mørup, M. & Schmidt, M.N. 2012. Bayesian Community Detection. in *Neural Computation 24 (9)*, 2434-2456.

Mørup, M., Madsen, K.H., Dogonowski, a.M., Siebner, H. & Hansen, L.K. 2010. Infinite Relational Modeling of Functional Connectivity in Resting State Fmri. in *Neural Information Processing Systems*, 1750-1758.

Ngo, D., Bellahsene, Z. & Todorov, K. 2013. Extended Tversky Similarity for Resolving Terminological Heterogeneities across Ontologies. in *OTM 2013, Lecture Notes in Computer Science, Vol.8185,* 711–718.

Pirrò, G. and Seco, N. 2008. Design, Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information Content. in *OTM Conferences (2)*, 1271–1288.

Pirrò, G. & Euzenat, J. 2010. a Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. in *International Semantic Web Conference (1),* 615–630.

Pitman J. 2002. Combinatorial Stochastic Processes. in *Tech. Rep., Springer*.

Schmidt, M.N., Herlau, T. & Mørup, M. 2012. Nonparametric Bayesian Models of Hierarchical Structure in Complex Networks. in *Informatics and Mathematical Modelling.*

Schmidt, M.N. & Mørup, M. 2013. Non-Parametric Bayesian Modeling of Complex Networks. an Introduction. in *IEEE Signal Processing Magazine, Vol. 30(3)*, 110-128

Schwartz, S. H. 2012. an Overview of the Schwartz Theory of Basic Values. in *Online Readings in Psychology and Culture, 2(1)*.

Stumme, G. & Mädche, a. 2001. Fca-Merge: Bottom-up Merging of Ontologies. in *Proceedings of 17th International Joint Conference on Artificial Intelligence (IJAI)*, 225–234

Wang, J., Wen, J.R., Lochovsky, F. & Ma, W.Y. 2004. Instance-based Schema Matching for Web Databases by Domain-Specific Query Probing. in *Proc. of the 30th Intl. Conference on Very Large Data Bases*, 408–419

Wasserman, S. & Anderson, C. 1987. Stochastic a Posteriori Blockmodels: Construction and Assessment. in *Sociological Networks, 9*, 1-36.

WVS: World Value Survey Association 2009. World Value Survey Official Data File V. 20090901. *Aggregate File Producer: ASEP/JDS, Madrid*

Xu, Z., Tresp, V., Yu, K., & Kriegel, H.P. 2006. Learning Infinite Hidden Relational Models. in *Uncertainty in Artificial Intelligence (UAI2006)*