2014 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 21–24, 2014, REIMS, FRANCE

NONPARAMETRIC STATISTICAL STRUCTURING OF KNOWLEDGE SYSTEMS USING BINARY FEATURE MATCHES

Morten Mørup[1], *Fumiko Kano Glückstad*[2], *Tue Herlau*[1], *Mikkel N. Schmidt*[1]

[1]Technical University of Denmark, Lyngby, Denmark, {mmor,tuhe,mnsc}@dtu.dk
[2]Copenhagen Business School, Frederiksberg, Denmark, fkg.ibc@cbs.dk

ABSTRACT

Structuring knowledge systems with binary features is often based on imposing a similarity measure and clustering objects according to this similarity. Unfortunately, such analyses can be heavily influenced by the choice of similarity measure. Furthermore, it is unclear at which level clusters have statistical support and how this approach generalizes to the structuring and alignment of knowledge systems. We propose a non-parametric Bayesian generative model for structuring binary feature data that does not depend on a specific choice of similarity measure. We jointly model all combinations of binary matches and structure the data into groups at the level in which they have statistical support. The model naturally extends to structuring and aligning an arbitrary number of systems. We analyze three datasets on educational concepts and their features and demonstrate how the proposed model can both be used to structure each system separately or to jointly align two or more systems. The proposed method forms a promising new framework for the statistical modeling and alignment of structure across an arbitrary number of systems.

***Index Terms*—** Bayesian non-parametrics, relational modeling, binary similarity, knowledge structuring.

1. INTRODUCTION

The representation, structuring, and alignment of domain knowledge is an important challenge within computer science, information science, cognitive science, computational linguistics and philosophy. Knowledge representation and alignment is needed for diverse applications such as the alignment of library classification systems [1], biomedical ontologies [2], and industry specific taxonomies [3]. In this work, we study the case where objects are represented by a number of binary features.

The structuring of knowledge within or across systems often relies on the computation of similarities between objects [4, 1, 2, 5]. One of the traditional approaches for computing similarities is the common extension comparison between two objects [4]. Let f_{xy} denote the number of features where object X takes the value $x \in \{0,1\}$ and object Y takes the value $y \in \{0,1\}$. A multitude of similarity measures have been employed for knowledge alignment based on different variations of quantifying similarity using these matches including the Jaccard index [6], Simple Matching Coefficient (SMC), and Tversky's Ratio Model [7] that respectively are

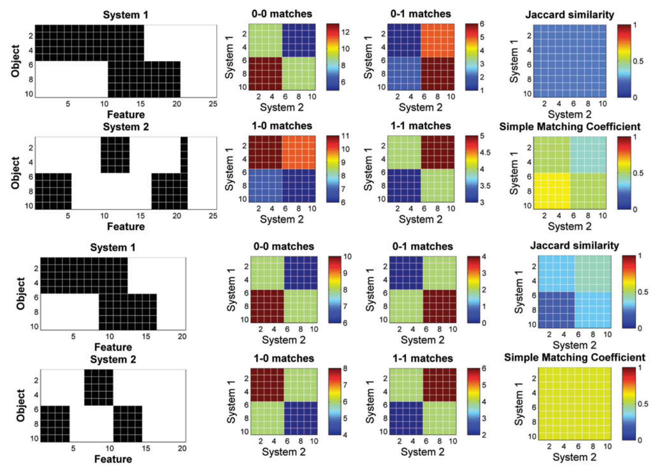


Fig. 1: Examples of similarity computation between features of two domains in which Jaccard (top) and SMC (bottom) fail in identifying the underlying (two) groups of objects in both domains despite the clear structures present in each of the different type of matches. By exploiting the structure of all types of matches (i.e., 0–0, 0–1, 1–0 and 1–1) the proposed framework for structuring binary similarity in both cases correctly identifies the existence of two groups in each system.

given by

$$\begin{aligned}\mathrm{Jaccard}(X,Y) &= f_{11}/(f_{11}+f_{10}+f_{01})\ ,\\ \mathrm{SMC}(X,Y) &= (f_{11}+f_{00})/(f_{11}+f_{10}+f_{01}+f_{00})\ ,\\ \mathrm{Tversky}(X,Y) &= f_{11}/(f_{11}+\upsilon f_{10}+\beta f_{01})\ .\end{aligned}$$

In Tversky's Ratio Model, υ and β are used to define the relative influence of 1–0 and 0–1 matches such that for $\upsilon=\beta=1$ Tversky's Ratio Model is equivalent to the Jaccard similarity. Unfortunately, the alignment of knowledge systems can be strongly affected by the similarity measure employed. Consider for instance the two examples given in Figure 1 in which the Jaccard index and SMC each are unable to detect structure between two groups of entities in two systems despite the clear structure in the raw f_{xy} matches.

Similarity computations between objects play important roles in the ontology alignment discipline, integrated with the structural information of ontologies. For example, performances of aligning industry specific taxonomies and regulations employing a similarity measure (i.e. cosine similarity, Jaccard similarity, or market basket analysis) combined with hierarchical structure information were compared in [3]. Performances of biomedical ontology mapping with a new ontological similarity measure, named Tversky's parameterized ratio model of similarity were investigated in [2] and compared against other measures derived from the idea of Tversky's Ratio model [8, 9]. While the performance evaluation of ontology matching is one branch of knowledge alignment, performance of similarity-based alignment and (re-)structuring of two knowledge systems was investigated in [10].

When structuring systems based on similarity an important open problem is to quantify the number of groups which are significant: Similarities based on few features are inherently more uncertain compared to similarity measures defined using many features. Consequently an alignment approach should take this uncertainty into account. We presently propose a non-parametric Bayesian generative model for the structuring of binary data that is not reduced to a specific projection of the different f_{xy} matches based on existing similarity measures such as Jaccard, SMC or Tversky's Ratio Model but exploits the entire structure of all matches simultaneously when structuring and aligning systems. The method can be used to simultaneously align an arbitrary number of systems and is able to automatically infer the number of groups within each system by quantifying the level in which groups have statistical support as defined by their structured interactions across the considered matches. The model proposed extends the existing block-modeling frameworks [11, 12, 13, 14] to the modeling of binary similarity matches and we demonstrate the models utility for the alignment of concepts within and between three educational systems.

2. MODEL FOR BINARY FEATURE MATCHES

We consider organizing knowledge both within and between systems. Let $f_c(i,j)$ be the number of matches between observation i and j of type $c \in \mathcal{C}$. For example, for binary data between two systems we have $\mathcal{C} = \{(00),(01),(10),(11)\}$. Let $\mathbf{f}(i,j)$ denote the vector of match counts for all types, and let $\boldsymbol{\mathcal{F}}$ denote the counts for all types and pairs of objects. Let further $N_{i,j} = \sum_c f_c(i,j)$ denote the total number of binary features.

A single system: We will initially consider structuring binary data from a single system such that $f_c(i,j) = \sum_n \delta[x_{in}, c(1)]\delta[x_{jn}, c(2)]$ where x_{in} denotes the nth feature of the ith object. As prior distribution for the clustering of objects to groups, $\boldsymbol{z}$, we will use the Chinese Restaurant Process (CRP) which defines a probability distribution over a partitioning of a set of objects. The CRP is parameterized by a concentration parameter α which governs its propensity to forming new groups (see also [15, 16]). Once the partitioning is formed we model the relative probability of observing each of the different types of matches between groups a and b, denoted $\boldsymbol{\eta}(a,b)$, by a Dirichlet prior. The prior is parameterized by $\boldsymbol{\eta}_0$ that specifies the relative extend to which the different types of matches occur. Finally, we model the vector of the number of matches of all types between object i and j, $\mathbf{f}(i,j)$, by a multinomial distribution parameterized by $\boldsymbol{\eta}(z_i,z_j)$ (where z_i denotes the group to which observation i is assigned) as well as the total number of features in the similarity computation $N(i,j)$. Thus, our generative model for organizing knowledge within one system based on all (four) types of binary similarity matches is given by

$$\begin{aligned}
\mathbf{z} &\sim \mathrm{CRP}(\alpha), && \text{objects to groups,}\\
\boldsymbol{\eta}(a,b) &\sim \mathrm{Dirichlet}(\boldsymbol{\eta}_0), && \text{proportion of types,}\\
\mathbf{f}(i,j) &\sim \mathrm{Mult}(\boldsymbol{\eta}(z_i,z_j), N(i,j)), && \text{no. of matches of each type.}
\end{aligned}$$

From the above generative model of $\boldsymbol{\mathcal{F}}$ the joint distribution can be derived and since the Dirichlet distribution is conjugate to the multinomial distribution $\boldsymbol{\eta}(a,b)$ can be analytically marginalized,

$$p(\boldsymbol{\mathcal{F}},\mathbf{z}|\alpha,\boldsymbol{\eta}_0) = \int \prod_{i\neq j} p(\mathbf{f}(i,j)|\mathbf{z},\boldsymbol{\eta})p(\boldsymbol{\eta}|\boldsymbol{\eta}_0)p(\mathbf{z}|\alpha)d\boldsymbol{\eta} = \left[\prod_{i\neq j}\frac{N(i,j)!}{\prod_c f_c(i,j)!}\prod_{a,b}\frac{\mathrm{B}\big(\mathbf{m}(a,b)\big)}{\mathrm{B}(\boldsymbol{\eta}_0)}\right]\cdot\frac{\Gamma(\alpha)\alpha^K}{\Gamma(\alpha+I)}\prod_k \Gamma(n_k),$$

where n_k is the number of objects in group k, I is the total number of objects, $\boldsymbol{m}(a,b) = \sum_{i\neq j}\boldsymbol{f}(i,j)\delta(z_i,a)\delta(z_j,b)$ is the number of matches between groups a and b, and $\mathrm{B}(\boldsymbol{\gamma}) = \frac{\prod_a \Gamma(\gamma_a)}{\Gamma(\sum_a \gamma_a)}$ is the multinomial Beta function.

Multiple systems: For structuring knowledge across multiple systems we extend the above model to consider all 2^D types of binary matches where $D \geq 2$ is the number of systems. For example, for three systems we have $\mathcal{C} = \{(000),(001),\ldots,(111)\}$. We introduce system specific groups such that the observed matches of type c are given by the D'th order tensor $\mathcal{F}_c$ with elements $f_c(i_1,\ldots,i_D)$,

$$\begin{aligned}
\mathbf{z}^{(d)} &\sim \mathrm{CRP}(\alpha_d), \quad \text{for } d \in \{1,\ldots,D\},\\
\boldsymbol{\eta}(a_1,\ldots,a_D) &\sim \mathrm{Dirichlet}(\boldsymbol{\eta}_0),\\
\mathbf{f}(i_1,\ldots,i_D) &\sim \mathrm{Mult}\big(\boldsymbol{\eta}(z^{(1)}_{i_1},\ldots,z^{(D)}_{i_D}), N(i_1,\ldots,i_D)\big).
\end{aligned}$$

We thereby obtain the following joint distribution

$$p(\boldsymbol{\mathcal{F}},\mathbf{z}|\boldsymbol{\alpha},\boldsymbol{\eta}_0) = \prod_{i_1,\ldots,i_D}\frac{N(i_1,\ldots,i_D)!}{\prod_c f_c(i_1,\ldots i_D)!} \prod_{a_1,\ldots,a_D}\frac{\mathrm{B}\big(\mathbf{m}(a_1,\ldots,a_D)\big)}{\mathrm{B}(\boldsymbol{\eta}_0)}\cdot\prod_d\frac{\Gamma(\alpha_d)\alpha_d^{K_d}}{\Gamma(\alpha_d+I_d)}\prod_{k_d}\Gamma(n_{k_d}),$$

where the notation is similar to that in the one system model. When $D = 2$ the model reduces to co-clustering the four types of binary matches between two systems.

According to the above model for structuring single and multiple systems, objects are structured according to their interactions across groups constituting homogenous blocks of binary match counts. The above framework is thereby closely related to stochastic block models [11, 12] and non-parametric extensions such as the infinite relational model [13, 14]. Thus, the proposed model can be considered an extension of the existing block-modeling frameworks to multinomial distributed count statistics.

Model inference: In order to infer $\boldsymbol{z}$ we use Markov chain Monte Carlo (MCMC) based on Gibbs sampling with split-merge moves [17, 18]. When modeling D systems, the Gibbs sampler cycles through the observations of each of the D systems one at a time and the split-merge sampler proposes merge or split moves within each system. The hyper-parameters $\boldsymbol{\eta}_0$ and $\boldsymbol{\alpha}$ are inferred by imposing an improper uniform prior, transforming the variable to the log-domain, and using a Metropolis-Hastings random walk procedure based on a normal proposal distribution with standard deviation 0.1. In our experiments the sampler was run for 1000 iterations where each iteration constituted a Gibbs sweep and 10 split-merge moves for inferring $\mathbf{z}$ within each system, 100 random-walk moves for each α_d parameter, and 10 random walk moves for each parameter in $\boldsymbol{\eta}_0$.

Table 1: Mean normalized mutual information (NMI) between runs as well as mean number of extracted components (K). In parenthesis is given the uncertainty on the least significant digit computed as the standard deviation of the mean across the 10 runs.

	System 1		System 2		System 3	
	NMI	K	NMI	K	NMI	K
Korea	1.00(0)	11.0(0)	-	-		
Czech	0.97(1)	10.0(0)	-	-		
Japan	0.95(2)	12.5(2)	-	-		
Korea-Czech	1.00(0)	6.0(0)	0.97(1)	15.4(2)	-	
Korea-Japan	0.99(1)	9.2(1)	0.94(2)	13.8(1)	-	
Czech-Japan	0.94(1)	11.6(2)	0.85(4)	6.0(0)	-	
Kor.-Cz.-Jap.	1.00(0)	10.0(0)	0.99(0)	24.4(2)	1.0(0)	21.0(0)

3. ALIGNMENT OF UNESCO EDUCATIONAL SYSTEMS

To demonstrate and evaluate our proposed model, we analyze a UNESCO dataset of educational systems.

Data: The datasets employed represents educational systems of Republic of Korea (Korea), Czech Republic (Czech), and Japan, which are available from the UNESCO Institute of Statistics (UIS) website[1]. The datasets consist of 55, 60 and 54 educational concepts respectively in Korea, Czech and Japan, and their 127 standardized features pre-defined by UIS in accordance with International Standard Classification of Education (ISCED). Fig. 2.a overviews distributions of educational concepts (K1-K55, C1-C60, J1-J54) in the six ISCED levels starting from pre-primary to tertiary education. Features other than the ISCED levels are for example program orientations, entrance requirements, starting age, cumulative duration, qualifications etc. Some of these features can be seen in the "shared features" columns of Fig. 2.b. Based on these data we create binary concept-features matrices for the three systems.

Results: We analyzed each system separately, two systems at a time, and all three systems jointly. In our analysis we ran the sampler ten times initialized differently both in terms of initial number of components and the order in which the modes were updated. We extracted from each run the sample with highest value of the joint-distribution and evaluated the number of components as well as normalized mutual information (NMI) between the ten runs. A NMI of 1 would indicate perfect correspondences between the extracted groups across runs. These results are given in Table 1. From the table it can be seen that the highest likelihood solution was very reliably identified when analyzing all three systems jointly as well as Korea separately with NMI all above 0.99 whereas the structure inferred when analyzing Czech and Japan jointly was the least stable with NMI of 0.94 and 0.85 for the two systems respectively.

In the following we display representative results obtained using a single run. Fig. 3.a illustrates the single system modeling (i.e., unipartite clustering) of the three educational systems. The graphs overview the statistically extracted structures in all types of binary matches for each single system and especially demonstrate that the Korean system has clear structure with relatively uniform clusters. The similar phenomenon is observable from the results of the Japanese system although the structures are slightly blurred compared to the Korean structures. Fig. 3.a reveals that more features

[1]http://www.uis.unesco.org/education/ISCEDmappings/Pages/default.aspx

2-a: Educational concepts and their ISCED levels

	ISCED 0	ISCED 1	ISCED 2	ISCED 3	ISCED 4	ISCED 5	ISCED 6	NC
	Pre-primary	Primary	Lower secondary	Upper secondary	Post-secondary, Non-tertieary	Tertiary (1. stage)	Tertieary (2. stage)	
Korea	K1-2	K3-5	K6-10	K11-20		K21-50	K51-56	
Czech	C1-3	C4-8	C9-18	C19-29	C30-37	C38--48	C49	C50-60
Japan	J1-3	J4-5	J6-8	J9-29	J30-34	J35-49	J50-52	J53-54

Cluster ID	Korean clusters	
	11 clusters	**shared features**
KU1	K26, K28, K29, K30, K31, K32, K33, K34, K39, K40, K41, K48, K50	Bachelor, ISCED5, require:3A, cumulative duration:16 etc.
KU2	K35, K36, K37, K43, K44, K45, K46, K47	ISCED3, require:2, cumulative duration: 12, starting age:15 etc.
KU3	K15, K16, K17, K18, K19, K20	Master, ISCED5, require:5A(1st), cum. duration:18, orientation:G etc.
KU4	K51, K52, K53, K54, K55	PhD, ISCED6, require:5A(2nd) cumulative duration:21+
KU5	K6, K7, K8, K9, K10	ISCED2, starting age:12, require:1, cumulative duration:9
KU6	K11, K12, K13, K14	Master, ISCED5, require:5A(1st), cum. duration:18, orientation:V etc.
KU7	K22, K23, K24, K49	Associate degree, ISCED5, destination:B, cumulative duration:14
KU8	K3, K4, K5	ISCED1, starting age:6
KU9	K25, K27, K42	ISCED5
KU10	K1, K2	ISCED0, start age: 3-5
KU11	K21, K38	ISCED5, cumulative ducation: 14-15

Cluster ID	Czech clusters	
	10 clusters	**shared features**
CU1	C9, C12, C14, C15, C16, C17, C18, C19, C20, C27, C38	adults:N, part-time:N, certification (ISCED2/3, orientation:G/P)
CU2	C21, C23, C24, C25, C28, C29, C31, C32, C34, C36	adults:Y/N, part-time:Y/N, certification (ISCED3/4, orientation:G/V/P)
CU3	C1, C2, C3, C4, C5, C6, C7, C8, C50, C51	certification, adults:N, part-time:N (start age:3/6/15, ISCED0/1/N)
CU4	C52, C55, C56, C58, C59, C60	part-time
CU5	C33, C39, C40, C41, C43, C45	adults:Y/N, part-time:Y/N (ISCED4/5)
CU6	C11, C13, C22, C26, C30, C37	certification (ISCED2/3/4)
CU7	C35, C46, C48, C49	adults, part-time (ISCED4/5/6)
CU8	C42, C44, C47	ISCED5, orientation:A, adults, part-time
CU9	C53, C54, C57	certification
CU10	C10	ISCED2, destination:C, orientation:P

Cluster ID	Japanese clusters	
	12 clusters	**shared features**
JU1	J9, J12, J14, J18, J20, J22, J25, J26, J28	ISCED3, duration:3, start age:15, require:2, cumulative duration:12
JU2	J10, J11, J13, J15, J16, J19, J21, J23	ISCED3, duration:3+, start age:15, require:2, cumulative duration:12+
JU3	J38, J41, J42, J47, J48, J49	ISCED5, adults:N, part-time:N
JU4	J30, J31, J32, J33, J34	ISCED4, duration:1+, start age:18
JU5	J35, J37, J44, J45, J46	ISCED5
JU6	J17, J24, J27, J29	ISCED3, destination:C, orientation:V etc.
JU7	J36, J39, J40, J43	ISCED5, certification/diploma
JU8	J1, J2, J3	ISCED0, duration 1-3
JU9	J50, J51, J52	ISCED6, require:5A(2nd)
JU10	J6, J7, J8	ISCED2, destination:A, cumulative duration:9
JU11	J4, J5	ISCED1, cumulative duration:6
JU12	J53, J54	certification, duration.1+

2-b: Unipartite clustering assignments

Korean (x Czech)	
6 (x 16) clusters	**Unipartite elements**
K1, K2, K22, K23, K24, K25, K26, K27, K28, K29, K30, K31, K32, K33, K34, K35, K36, K37, K39, K40, K41, K42, K43, K44, K45, K46, K47, K48, K49, K50	Merger of KU1, KU2, KU7, KU9, KU10
K11, K12, K13, K14, K15, K16, K17, K18, K19, K20	Merger of KU3, KU6
K6, K7, K8, K9, K10	KU5
K51, K52, K53, K54, K55	KU4
K3, K4, K5	KU8
K21, K38	KU11
Czech (x Korea)	
16 (x 6) clusters	**Unipartite elements**
C11, C13, C21, C25, C28, C29, C31, C32, C36	part of CU6 and CU2
C9, C12, C14, C15, C16, C17, C18, C19, C20	part of CU1
C4, C5, C6, C7, C8, C51	part of CU3
C40, C41, C43, C45, C47, C48	part of CU5, CU7, CU8
C55, C56, C58, C59, C60	part of CU4
C1, C2, C3, C50	part of CU3
C53, C54, C57	CU9
C22, C23, C24	part of CU2, CU6
C10, C34, C39	CU10, part of CU2, CU5
C27, C30, C38	part of CU1, CU6
C26, C37	part of CU6
C42, C44	part of CU8
C33, C35	part of CU5, CU7
C52	part of CU4
C46	part of CU7
C49	part of CU7

Czech (x Japan)	
12 (x 6) clusters	**Unipartite elements**
C9, C12, C14, C15, C16, C17, C18, C19, C20, C27, C30, C37, C38	CU1 and part of CU6
C11, C13, C21, C25, C28, C29, C36	part of CU6, CU2
C40, C42, C43, C44, C45, C47, C48	CU8 and part of CU5, CU7
C4, C5, C6, C7, C8, C51	part of CU3
C52, C55, C56, C58, C59, C60	CU4
C31, C32, C34, C39, C41	part of CU2 and CU5
C1, C2, C3, C50	part of CU3
C22, C23, C24, C26	part of CU2and CU6
C53, C54, C57	CU9
C46, C49	part of CU7
C33, C35	part of CU5 and CU7
C10	CU10
Japan (x Czech)	
6 (x 12) clusters	**Unipartite elements**
J9, J10, J11, J12, J13, J14, J15, J16, J17, J18, J19, J20, J21, J22, J23, J24, J25, J26, J27, J28, J29	Merger of JU1, JU2, JU6
J38, J39, J40, J41, J42, J43, J47, J48, J49, J50, J51, J52	JU3, JU9, JU7 (J39, J40, J43)
J6, J7, J8, J30, J31, J32, J33, J34	Merger of JU4, JU10
J35, J37, J44, J45, J46	JU5
J1 J2 J3, J36	JU8, JU7 (J36)
J4, J5, J53, J54	JU11, JU12

2-c: Bipartite clustering assignments

Korean (x Japan)	
9 (x 14) clusters	**Unipartite elements**
K25, K26, K27, K28, K29, K30, K31, K32, K33, K34, K39, K40, K41, K42, K48, K50	Merger of KU1, KU9
K11, K12, K13, K14, K15, K16, K17, K18, K19, K20	Merger of KU3, KU6
K35, K36, K37, K43, K44, K45, K46, K47	KU2
K51, K52, K53, K54, K55	KU4
K6, K7, K8, K9, K10	KU5
K22, K23, K24, K49	KU7
K3, K4, K5	KU8
K1, K2	KU10
K21, K38	KU11
Japan (x Korean)	
14 (x 9) clusters	**Unipartite elements**
J9, J10, J12, J13, J14, J18, J19, J20, J21, J22, J25, J26	majority of JU1 and part of JU2
J30, J31, J32, J33, J34, J39, J40	JU4 and part of JU7
J11, J15, J16, J23, J29	part of JU2 and JU6
J4, J5, J53, J54	Merger of JU11, JU12
J17, J24, J27, J28	majority of JU6 and part of JU1
J35, J36, J37	part of JU5, JU7
J6, J7, J8	JU10
J1, J2, J3	JU8
J41, J42, J44	part of JU3, JU5
J50, J51., J52	JU9
J43, J49	part of JU7, JU3
J47, J48	part of JU3, JU5
J45, J46	part of JU5
J38	part of JU3, JU5

Fig. 2: **a:** Overview of distributions of educational concepts classified based on the pre-defined ISCED levels. **b:** The unipartite clustering solution partitions educational concepts in the respective systems into clusters and what features are commonly shared among the members of the obtained clusters. **c:** Members of clusters restructured by the bipartite clustering solution. The "unipartite elements" columns indicate how the members of the unipartite clusters are reorganized when two educational systems are aligned.

are shared among members of each cluster in the Korean system as red color dominates in 1–1 matches, but also identifies relations across extracted clusters (e.g. KU1 and KU9). The shared features listed in Fig. 2.b show that the partitions of both the Korean- and the Japanese systems are primarily influenced by the classification of the ISCED levels, while the Czech system is strongly influenced by other criteria, e.g. whether an educational concept is targeted for adult education/part-time education or not. Accordingly the extracted clusters in the Czech system consist of members belonging to several ISCED levels. In addition, the graph drawn by the Jaccard similarities within Czech system fails to identify some of the cluster structures observable in the graph of the 1–1 matches.

In Fig. 3.b, the bipartite clustering of our framework aligns two of the educational systems in three combinations (i.e. Korea-Czech, Korea-Japan, Czech-Japan). The graphs indicate that Korea-Japan

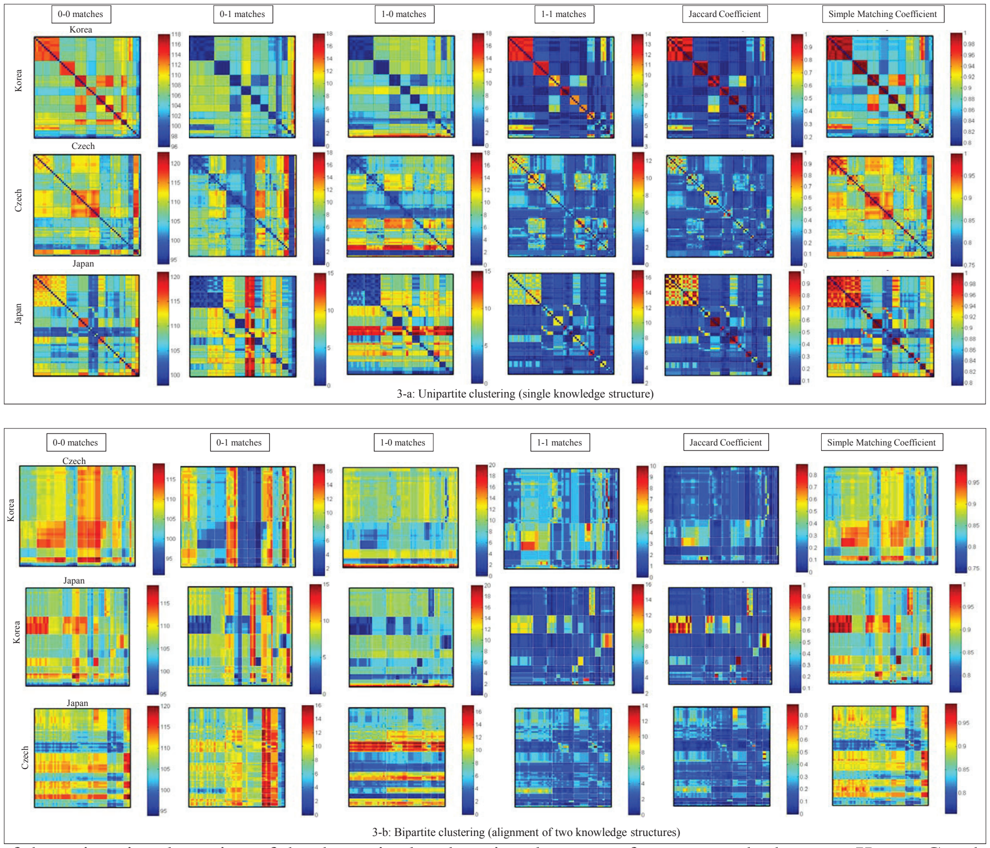
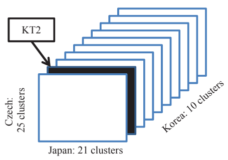


Fig. 3: **a:** Results of the unipartite clustering of the three single educational systems from top to the bottom: Korea; Czech; Japan. The graphs from the first to the fourth columns depict structures in each of the four types of binary matches (i.e., 0–0, 0–1, 1–0 and 1–1). The graphs of the fifth and the sixth columns are drawn based on similarity scores (Jaccard and SMC) between all combinations of educational concepts within each single system, which are reorganized according to the extracted clusters. The structures of the 1-1 and 0-0 matches are relatively identical to the Jaccard and SMC structures. Our framework additionally provides structural information supported by all types of binary matches. Members of the extracted clusters and features shared among the cluster members are listed in Fig. 2-b. **b:** Bipartite clusterings of two educational systems: Korea-Czech; Korea-Japan; Czech-Japan. The graphs indicate how the original structures in all types of binary matches of each single system influence on the alignment of the two systems.

bipartite clustering has clearer structures in 1-1 matches than the other two combinations, as the structures in the two systems are integrated based on e.g. the ISCED levels (the classification criteria influencing on the Korean- and the Japanese unipartite clusterings). Another important finding is that the influence of the original unipartite structures is revealed in the bipartite clusterings. Fig. 2.c shows how the original unipartite clusters are merged or split when the two systems are aligned. For example, when the Korean system is aligned with the Czech system, some of the unipartite Korean clusters (KUs) are merged into one cluster, otherwise the same KUs remain as they are identified by unipartite solution. On the other hand, the majority of the unipartite Czech clusters (CUs) are split into several smaller groups and some of these groups are merged into one as a new cluster in Fig. 2.c, influenced by the unipartite Korean clusters clearly partitioned by e.g the ISCED levels. This phenomenon is supported by the graphs of 1–0 and 0–1 matches in the Korea-Czech alignment in Fig. 3.b where the Czech specific feature structures (0–1 matches) have more fine-grained stronger structures than the Korean specific feature structures (1–0 matches) with coarse weaker structures.

As outlined in Fig. 5, the 3-way clustering here identifies 10 Korean clusters (KT1-10), 25 Czech clusters (CT1-25) and 21 Japanese clusters (JT1-21). As the Korean system has the clearer and simpler *unipartite* partitions according to the ISCED levels and has the coarse and weaker structures in the (1–0 matches) of the *bipartite* clusterings in Fig. 3.b, the number of KTs obtained are substantially fewer than the other two systems. The extracted clusters are fine-grained and generally well-separated, i.e. similar types of educational concepts are grouped together without irrelevant concepts. For example, the tables at the top of Fig. 4 lists feature vectors of cluster members indicating stronger interactions with the second Korean cluster (KT2). More specifically, the KT2 cluster has stronger interactions with CT11, CT13, CT21 as well as JT10, JT11, JT16 in "K-C-0" and "K-0-J" matches. The list of feature vectors demonstrate that the second Korean cluster (KT2) consists of educational concepts at the level of ISCED 5 providing Master degree and almost all eight members of KT2 possess uniform features. On the other hand, CT11, CT13, CT21 as well as JT10, JT11, JT16, all of which belong to the ISCED5 level, are divided into several sub-clusters based on different classification dimensions such as *part-time*, *adult*, *degree* etc. This implies that the KT2 is aligned with several CTs in the "K-C-0" matches and with several JTs in the "K-0-J" matches. In this



KT2	K35	ISCED:5	destination:B	ISCED5:Long	degree:2nd	ISCEDrequirement:5A(1st)	tertiary:Master	Master	startingage:22	duration:2	cumulativeduration:18	workbased:No	adults:No	parttime:No
	K36	ISCED:5	destination:A	ISCED5:Long	degree:2nd	ISCEDrequirement:5A(1st)	tertiary:Master	Master	startingage:22	duration:2	cumulativeduration:18	workbased:No	adults:No	parttime:No
	K37	ISCED:5	destination:A	ISCED5:Long	degree:2nd	ISCEDrequirement:5A(1st)	tertiary:Master	Master	startingage:22	duration:2	cumulativeduration:18	workbased:No	adults:No	parttime:No
	K43	ISCED:5	destination:A	ISCED5:Long	degree:2nd	ISCEDrequirement:5A(1st)	tertiary:Master	Master	startingage:22	duration:2	cumulativeduration:18	workbased:No	adults:No	parttime:No
	K44	ISCED:5	destination:A	ISCED5:Long	degree:2nd	ISCEDrequirement:5A(1st)	tertiary:Master	Master	startingage:22	duration:2	cumulativeduration:18	workbased:No	adults:No	parttime:No
	K45	ISCED:5	destination:A	ISCED5:Long	degree:2nd	ISCEDrequirement:5A(1st)	tertiary:Master	Master	startingage:22	duration:2	cumulativeduration:18	workbased:No	adults:No	parttime:No
	K46	ISCED:5	destination:A	ISCED5:Long	degree:2nd	ISCEDrequirement:5A(1st)	tertiary:Master	Master	startingage:22	duration:2	cumulativeduration:18	workbased:No	adults:No	parttime:No
	K47	ISCED:5	destination:A	ISCED5:Long	degree:2nd	ISCEDrequirement:5A(1st)	tertiary:Master	Master	startingage:22	duration:2	cumulativeduration:18	workbased:No	adults:No	parttime:No
CT11	C47	ISCED:5	destination:A	ISCED5:Long/verylong	degree:2rd/3rd +	ISCEDrequirement:5A		diploma	startingage:24+	duration:1+	cumulativeduration:19+	workbased:No	adults:yes	parttime:yes
	C48	ISCED:5	destination:A	ISCED5:Long/verylong	degree:2nd/3rd	ISCEDrequirement:5A(2nd)		JurisDoctor	startingage:24	duration:1	cumulativeduration:19	workbased:No	adults:Yes	parttime:Yes
CT13	C43	ISCED:5	destination:A	ISCED5:Long	degree:1st	ISCEDrequirement:3A,4A		Bachelor	startingage:19	duration:5	cumulativeduration:18	workbased:No	adults:Yes/No	parttime:Yes/No
	C45	ISCED:5	destination:A	ISCED5:Long	degree:1st	ISCEDrequirement:3A,4A		Doctor	startingage:19	duration:6	cumulativeduration:19	workbased:No	adults:Yes/No	parttime:Yes/No
CT21	C44	ISCED:5	destination:A	ISCED5:Long	degree:2nd	ISCEDrequirement:5A(1st)	tertiary:Master	Bachelor	startingage:22	duration:2–3	cumulativeduration:18	workbased:No	adults:Yes/No	parttime:Yes/No
JT10	J47	ISCED:5	destination:A	ISCED5:Long	degree:2nd	ISCEDrequirement:5A(1st)	tertiary:Master	Master	startingage:22	duration:2	cumulativeduration:18	workbased:No	adults:No	parttime:No
	J48	ISCED:5	destination:A	ISCED5:Long	degree:2nd	ISCEDrequirement:5A(1st)	tertiary:Master	Master	startingage:22	duration:2	cumulativeduration:18	workbased:No	adults:No	parttime:No
JT11	J45	ISCED:5	destination:A	ISCED5:Long	degree:2nd	ISCEDrequirement:5A(1st)	tertiary:Master	Master	startingage:22	duration:2+	cumulativeduration:18+	workbased:No	adults:No	parttime:Yes
	J46	ISCED:5	destination:A	ISCED5:Long	degree:2nd	ISCEDrequirement:5A(1st)	tertiary:Master	Master	startingage:22	duration:2+	cumulativeduration:18+	workbased:No	adults:No	parttime:Yes
JT16	J43	ISCED:5	destination:A	ISCED5:Long	degree:Intermediate	ISCEDrequirement:5A(1st)		Certification	startingage:22	duration:1+	cumulativeduration:17+	workbased:No	adults:No	parttime:No
	J49	ISCED:5	destination:A	ISCED5:Long	degree:2nd	ISCEDrequirement:5A(1st)	tertiary:Master	JurisDoctor	startingage:22	duration:3	cumulativeduration:19	workbased:No	adults:No	parttime:No

Fig. 4: Feature vectors of cluster members that indicate stronger interactions with the second Korean cluster (KT2).

way, the one-to-many cluster relations between KT vs. CTs or JTs are extracted. Among these interactions, KT2, CT21 and JT10/JT11 are the interactions where all three systems simultaneously interact with each other (K-C-J matches). To sum up, the 3-way clustering jointly models all 8 types of binary matches and identifies structures of interactions among the obtained clusters within and across the three systems.

4. CONCLUDING REMARKS AND FUTURE WORKS

Alignment and structuring of one or more knowledge systems usually requires similarity computations between objects belonging to the respective knowledge systems. As pointed out in [10, 19], the selection of a similarity measure influences heavily on the performance of the knowledge alignment and (re-)structuring when employing a co-clustering algorithm based on dense relational modeling [20]. The recent trend of ontological similarity measures extends Tversky's Ratio model by integrating fuzzy set representations and information content of a concept based on information theory [2]. These types of new similarity measures are theoretically relevant to the alignment and structuring of knowledge. A question is how our proposed framework is positioned in contrast to such recent progress in knowledge alignment and structuring. One argument is that our proposed framework enables the visual inspection of the overall feature structures while aligning and extracting knowledge structures at the level of statistical support. Perhaps, the most unique noteworthy argument is that our framework enables the simultaneous alignment of three or more knowledge systems since the framework is not restricted to the similarity computation between two objects. The contrastive analysis of the unipartite, bipartite and tripartite clusterings demonstrated that the rather abstract cluster like KT2 in the uniformly structured Korean system is aligned with several fine-grained clusters (i.e. subsumed clusters) in the opposed systems. This indicates that our framework enables the identification of latent ontological structures within and across the systems. Thus, one of our future challenges is to develop hierarchical ontologies linked within and across the three systems to facilitate interpretation.

Another issue to be considered is that we have in our experiment employed small datasets organized by taxonomic information. A question is the applicability of our framework to different types of datasets, i.e. from strictly structured ontologies to sparse and unstructured large datasets where model inference will be substantially more challenging. In addition, we need to investigate objective evaluation criteria of the results, which has been considered outside the scope of this paper. A noteworthy remark is that our unipartite solution enables to assess whether an input dataset is structured or unstructured as shown in our experiment. This information itself is useful when two or more datasets are to be aligned afterwards. Such additional functionality can also be assessed when testing the applicability of our framework with different types of datasets.

5. REFERENCES

[1] A. Isaac, L. V. D. Meij, S. Schlobach, and S. Wang, "An empirical study of instance-based ontology matching," *The Semantic Web, Lecture Notes in Computer Science*, vol. vol4825, pp. 253–266, 2007.

[2] V. Cross, X. Yu, and X. Hu, "Unifying ontological similarity measures: A theoretical and empirical investigation," *International Journal of Approximate Reasoning*, vol. vol54, pp. 861–875, 2013.

[3] C. P. Cheng, G. T. Lau, K. H. Law, J. Pan, and A. Jones, "Regulation retrivial using industry specific taxonomies," in *AI and Law*, vol. 16, pp. 277–303. 2008.

[4] J. Euzenat and P. Shvaiko, *Ontology Matching*, Springer, 2007.

[5] D. Ngo, Z. Bellahsene, and K. Todorov, "Extended tversky similarity for resolving terminological heterogeneities across ontologies," *OTM 2013, Lecture Notes in Computer Science*, vol. vol8185, pp. 711–718, 2013.

[6] P. Jaccard, "Distribution de la flore alpine dans le bassin des dranses et dans quelques regions voisines," in *Bulletin de la societe vaudoise des sciences naturelles,vol.37*, pp. 241–272. 1901.

[7] A. Tversky, "Features of similarity," *Psychological Review*, , no. 4, pp. 327–352, 1977.

[8] G. Pirrò and N. Seco, "Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content," in *OTM Conferences (2)*, 2008, pp. 1271–1288.

[9] Pirrò G. and J. Euzenat, "A feature and information theoretic framework for semantic similarity and relatedness," in *International Semantic Web Conference (1)*, 2010, pp. 615–630.

[10] F. K. Glückstad, T. Herlau, M. N Schmidt, and M. Mørup, "Cross-categorization of legal concepts across boundaries of legal systems: in consideration of inferential links," in *AI and Law*. Springer, 2013.

[11] S. Wasserman and C. Anderson, "Stochastic a posteriori blockmodels: Construction and assessment," 1987.

[12] K. Faust and S. Wasserman, "Blockmodels: Interpretation and evaluation," 1992.

[13] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, "Learning systems of concepts with an infinite relational model," in *AAAI*, 2006, vol. 3, p. 5.

[14] Z. Xu, V. Tresp, K. Yu, and H.P. Kriegel, "Learning infinite hidden relational models," *Uncertainty in Artificial Intelligence (UAI2006)*, 2006.

[15] D. Aldous, "Exchangeability and related topics," *École d'Été de Probabilités de Saint-Flour XIII1983*, pp. 1–198, 1985.

[16] J. Pitman et al., "Combinatorial stochastic processes," Tech. Rep., Springer, 2002.

[17] S. Jain and Radford M Neal, "A split-merge markov chain monte carlo procedure for the dirichlet process mixture model," *Journal of Computational and Graphical Statistics*, vol. 13, no. 1, 2004.

[18] D.B. Dahl, "Sequentially-allocated merge-split sampler for conjugate and nonconjugate dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, vol. 11, 2005.

[19] F. K. Glückstad, T. Herlau, M. N Schmidt, and M. Mørup, "Unsupervised knowledge structuring: Application of infinite relational models to the fca visualization," in *Proc. 9th International Conference on Signal Image Technology and Internet Based Systems (SITIS 2013), Kyoto, Japan*, 2013.

[20] T. Herlau, M. Mørup, M. N. Schmidt, and L. K. Hansen, "Modelling dense relational data," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Santander, Spain*. 2012.

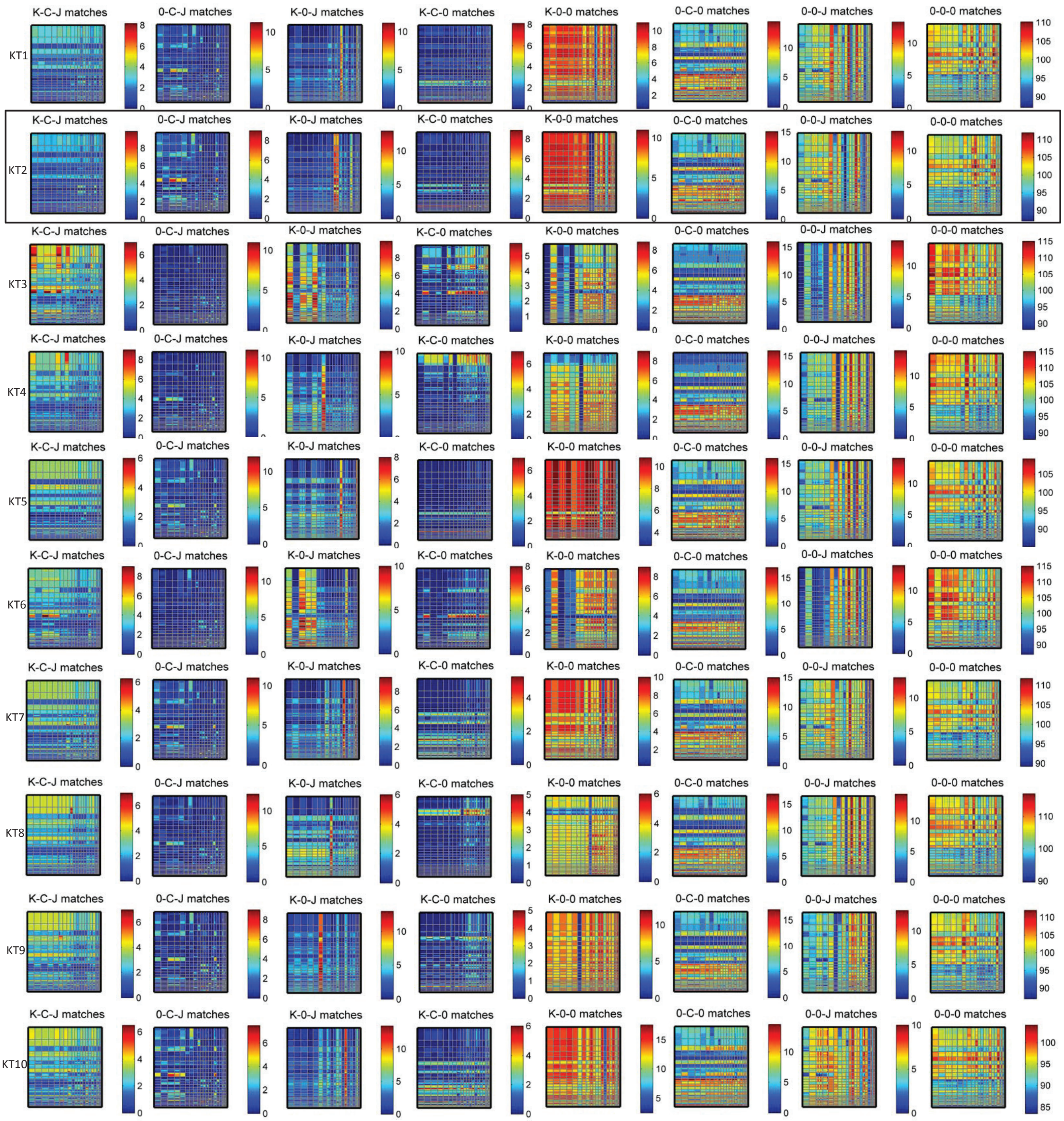


Fig. 5: 3-way clustering of the three educational systems (Korea-Czech-Japan). The graphs overview how the 10 Korean clusters (the number of matches are averaged among members of each cluster) interact with the other two systems. The 10 rows from the top to the bottom refer to the Korean clusters from KT1 to KT10, and the 8 columns from the left to the right refer to 111, 011, 101, 110, 100, 010, 001, and 000 matches. For convenience, these combinations are denoted as [KCJ, 0CJ, K0J, KC0, K00, 0C0, 00J, 000] in order to specify which systems interact with each other.