1K1-IOS-1a-4

Application of the Infinite Relational Model combined with the Bayesian Model of Generalization for Effective Cross-Cultural Knowledge Transfer

Fumiko Kano Glückstad [*1]

[*1] Dept. of International Business Communication
Copenhagen Business School (CBS)

Morten Mørup [*2]

[*2] Section for Cognitive Systems
DTU Informatics
Technical University of Denmark

This paper investigates how the Infinite Relational Model (IRM) [Kemp 2006], a novel unsupervised machine learning method, is effectively applied to loosely-structured datasets consisting of concepts and features for the purpose of mapping Culturally Specific Concepts (CSCs) in a multi-cultural context. The aim of this investigation is two-fold: **i)** to identify an effective strategy of applying the IRM for the purpose of CSC-mapping; and **ii)** to investigate possibilities of applying the IRM for efficiently constructing feature-based ontologies that are multi-culturally interoperable. Accordingly, three strategies are tested in our experiments: **1)** applying the IRM directly to two CSC-feature matrices, respectively representing the educational domain knowledge in Japan and Denmark for first categorizing them into categorical classes that are to be subsequently compared and aligned; **2)** applying the IRM directly to a matrix where the two CSC-feature matrices respectively representing the Danish- and Japanese educational domain knowledge are merged; and **3)** applying the Bayesian Model of Generalization (BMG) [Tenenbaum 2001] to directly compute similarity relations between CSCs in the two cultures at hand, thereafter to apply the IRM for clustering CSCs in the respective cultures into categorical classes. The results indicate that the third strategy seems to be the most effective approach for not only clustering CSCs into more specific and appropriate categorical classes but also for capturing complex relationships between each categorical classes existing in the two cultures.

1. Introduction

The recent internet revolution with its fast-paced globalization has brought about new possibilities for people located thousands of miles apart to real-time communicate with each other. Although we mostly use English as a common communication code, misunderstandings are almost unavoidable in contemporary cross-cultural communications. This implies that multilinguality is inherently challenged to effectively support human perception of concepts existing in diverse socio-cultural communities. Within the ontology research domain, there are several large-scale frameworks that link multi-cultural information in a complex manner. [Cimiano, 2011] compares these multilingual ontology frameworks such as the KYOTO project [Vossen, 2008] and the MONNET project [Declerck, 2010] based on a number of dimensions used in categorizing different types of ontology localization projects [Espinoza, 2009]. These dimensions are: *international (standardized) vs. culturally influenced domains; functional vs. documental localization; and interoperable vs. independent ontology* [Cimiano, 2011].

In [Glückstad, 2012-a], [Glückstad, 2012-b], potentially applicable feature-based similarity measures that can be used for mapping *independent ontologies* in a *culturally influenced domain* in a *functional* manner are compared based on qualitative analyses. These analyses identified that the Bayesian Model of Generalization (BMG) [Tenenbaum 2001] is the most intuitive and effective measure of mapping CSCs existing in two cultures. The application of the BMG requires highly appropriate datasets consisting of CSCs and their definitional features. In [Glückstad, 2012-a], an empirical study was performed with datasets obtained from a semi-automatic feature-based ontology construction method known as Terminological Ontology (TO) proposed by [Madsen 2004]. The results from that study indicated that particularly strict rules for constructing TOs may risk causing the elimination of important features. It means that the original TO-approach may require a more flexible taxonomic organization of feature structures.

Inspired by the previous works, we here investigate how the Infinite Relational Model (IRM) [Kemp 2006], a novel unsupervised machine learning method, is combined with the BMG for efficiently mapping CSCs and is applied for constructing more flexible feature structures of taxonomies. Both the IRM and the BMG applied in this work are originally proposed by cognitive scientists [Kemp, 2006] and [Tenenbaum, 2001]. While [Kemp, 2006] emphasizes that the IRM considers the semantic knowledge problem from the viewpoint of: *how representations of semantic knowledge are acquired*, [Tenenbaum, 2001] addresses three crucial questions of learning raised by [Chomsky, 1986]: **1)** *What constitutes the learner's knowledge?*; **2)** *How does the learner use that knowledge to decide how to generalize;* and **3)** *How can the learner acquire that knowledge from the example encountered?* Thus, the combination of the IRM and the BMG is potentially an interesting attempt from a philosophical point of view. However, in this paper we focus on how the IRM and the BMG are efficiently combined from a practical point of view, and the philosophical and pragmatic discussions are dealt with in a subsequent research report [Glückstad, 2012-c].

Relevant works that apply the BMG to a practical problem of mapping ontologies do not exist to our knowledge. However,

Contact: Fumiko Kano Glückstad, CBS, DK-2000 Frederiksberg, Denmark, E-mail: {fkg.ibc | fkg.isv}@cbs.dk

Tversky's set-theoretic model [Tversky, 1977] on which the BMG is based, is widely known in the area of ontology matching such as shown in [Huang, 2010] and [de Souza, 2004]. The IRM has been applied to diverse research domains among others in the area of neuroimaging where functional groups and their interactions are extracted by the IRM [Mørup, 2010] and in the area of collaborative filtering and topic modeling [Xu, 2006], [Hansen, 2011].

In the next section, the experimental settings and strategies employed in this work are explained in detail, followed by brief reviews of the BMG and the IRM in Section 3. Section 4 analyzes the results obtained from the three experimental strategies. In Section 5, we discuss some critical issues and future perspectives, followed by conclusions in Section 6.

2. Experimental settings

2.1 Data source

In this work, we first create two datasets, respectively representing the Danish educational domain knowledge and the Japanese educational domain knowledge. The datasets consists of educational terms and their definitional features that are manually extracted from text corpora. The Japanese corpora used for this experiment are: **1)** "Outline of the Japanese School System" published by the Center for Research on International Cooperation in Educational Development (CRICED), University of Tsukuba, Japan; and **2)** "Higher Education in Japan" published by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT). The Danish documents are downloaded from the Euridice web-site published by the Education, Audiovisual and Culture Executive Agency under the EU commission. These corpora are written in English and hence it is feasible to identify original expressions of educational terms in the respective languages from existing parallel- or content aligned corpora. This enables us to eventually achieve translation between Japanese CSCs and Danish CSCs through the English term mapping.

The CSCs and their definitions, all written in English, are manually extracted from the text corpora, e.g. the Danish CSC "municipal school (DA: *folkeskole*)" and its definition "*a comprehensive school covering both primary and lower secondary education, i.e. one year of pre-school class, the first (grade 1 to 6) and second (grade 7-9/10) stage basic education, or in other words it caters for the 6-16/17-year-olds*". From this definition, we create a feature list consisting of "comprehensive school" "primary and lower secondary education" "basic education" "targeted for 6-16/17 years old". This definition also implies that "municipal school" is categorized into three sub-CSCs "preschool class", "first stage" and "second stage", respectively having their features "one year preschool education" "1-6 grades" and "7-9/10 grades". These sub-CSCs are supposed to inherit features defined in the subordinate CSC, in this case "municipal school". In this way, 59 Danish CSCs and 54 Japanese CSCs and their features all written in English are listed up. In addition, some features are manually standardized, e.g. a feature "continuing education for adults" in Denmark is standardized with a feature "opportunities for life-long learning" in Japan. Finally, the following operations are manually implemented: **a)** If a feature value in one country is completely included in a feature value in the other country (e.g. the feature "6-12 y.o." in Japan is completely included in the feature "6-17 y.o." in Denmark), a CSC possessing the feature that includes the other feature (a CSC possessing "6-17 y.o." should also possess "6-12 y.o."; and **b)** If two features from the respective countries are partly overlapping (e.g. "13-15 y.o." in Japan and "14-17 y.o." in Denmark are partly overlapping), a dummy feature referring to the exact overlapping range (e.g. "14-15 y.o.") is created. In this example, a Japanese CSC possessing "13-15 y.o." should also possess the dummy feature "14-15 y.o.". In the same way, a Danish CSC possessing "14-17 y.o." should also possess the dummy feature "14-15 y.o.".

Accordingly, in total 229 features are registered in the two matrices, respectively representing the Danish- and Japanese educational systems. In each matrix, if a feature is possessed by a CSC, the numeric value "1" appears, otherwise "0" is assigned. In these matrices, the Danish- and Japanese CSCs are respectively denoted as D_i and J_j and feature IDs are assigned as f_k. Both the Danish- and Japanese CSCs and their features are alphabetically registered. Hence, it requires systematic taxonomic organization of CSCs for achieving effective CSC-mapping between the two cultures.

2.2 Experimental strategies

[Kemp 2006] emphasizes that the IRM considers the semantic knowledge problem from the viewpoint of: *how representations of semantic knowledge are acquired*, instead of starting from the viewpoint of how these systems can be represented. This has inspired us to investigate how semantic knowledge is acquired from the limited domain text corpora and how this can effectively be represented for the purpose of cross-cultural knowledge transfer. Accordingly, three strategies are tested in the experiments: **1)** applying the IRM directly to the respective CSC-feature matrices for first categorizing them into categorical classes that are to be subsequently compared and aligned; **2)** applying the IRM directly to a matrix where the two CSC-feature matrices, respectively representing the Danish- and Japanese educational domain knowledge are merged; and **3)** applying the BMG to directly compute similarity relations between CSCs in the two cultures, and thereafter to apply the IRM for clustering CSCs in the respective cultures into categorical classes. This implies that, in case of strategy **1)** and **2)**, the IRM clusters CSCs based on CSC-feature links, whereas, in case of strategy **3)**, the clustering is based on CSC-CSC links between Danish and Japanese CSCs which are identified by the BMG. These three strategies are compared in Section 4 and 5.

3. Methods

3.1 The Bayesian Model of Generalization (BMG)

The BMG [Tenenbaum 2001] is a cognitive model, which uniquely unifies the two following classically opposing approaches to similarity and generalization: Tversky's *set theoretic model of similarity* [Tversky 1977] and Shepard´s *continuous metric space model of similarity and generalization* [Shepard 1987]. The philosophical background of the BMG described in [Tenenbaum 2001] is reviewed in [Glückstad, 2012-

c]. A key point in the BMG is to compute *the conditional probability that y falls under C (Consequential Region) given the observation of the example x* based on the following formula:

$$P\ (y \in C|x) = 1/[1 + \frac{\sum_{h:x\in h, y\notin h} p(h,x)}{\sum_{h:x,y\in h} p(h,x)}] \quad (1)$$

The consequential region *C* in our work indicates the categorical region where a new object *y* belongs. In equation (1), a hypothesized subset *h* is defined as the region where a concept belongs to *h*, if and only if, it possesses feature *k* [Tenenbaum 2001]. It means, for our work, that *y* is considered as a newly encountered object existing in a Source Culture (SC) domain when *y* is introduced to a Target Culture (TC) audience and the TC audience compares this new object *y* with an observed data *x* which is part of his/her prior knowledge (referent dataset).

Another unique point of the BMG is that $P(h, x) = P(x|h)P(h)$ in equation (1), represents the weight assigned to the consequential subset *h* in terms of the example *x*. This can be achieved by specifically assigning the weight $P(h, x)$ based on the strong sampling scheme defined in [Tenenbaum 2001] as follows:

$$P(x|h) = \begin{cases} 1/|h| & if\ x \in h \\ 0 & otherwise \end{cases} \quad (2)$$

Here, $|h|$ indicates the size of the region *h* [Tenenbaum 2001]. In our work, the number of objects possessing the k^{th} feature in the referent dataset is considered as the size of the region *h*. [Tenenbaum 2001] explains that the prior $P(h)$ is not constrained in their analysis so that it can accommodate arbitrary flexibility across contexts. Hence in this work, $P(h) = 1$.

3.2 The Infinite Relational Model (IRM)

According to [Kemp, 2006], a key feature of the IRM is to *automatically choose an appropriate number of clusters using a prior that favors small numbers of clusters, but has access to a countably infinite collection of clusters*. In [Kemp 2006], the observed data are considered as *m* relations involving *n* types. For the experimental strategies **1)**, **2)** and **3)** in our work, we apply the simplest model: dealing with two types with a single two-place relation *R*: T_1 x $T_2 \rightarrow \{0, 1\}$. More specifically, in strategies **1)** and **2)** T_1 corresponds to either Danish and/or Japanese CSCs and T_2 corresponds to definitional features, while in strategy **3)**, T_1 and T_2 respectively corresponds to Danish CSCs and Japanese CSCs.

The principle of generating clusters in the IRM, according to [Kemp, 2006], is based on a distribution over partitions induced by a so-called Chinese Restaurant Process (CRP) [Pitman, 2002]. The CRP starts a partition process with a single cluster containing a single object. The i^{th} object has possibilities to belong to either of the following:

- A new cluster with probability: γ / (i-1+ γ)
- An existing cluster with probability: n_a / (i-1+ γ)

Here, n_a is the number of objects already assigned to cluster *a*, and γ is a parameter [Kemp, 2006]. The CRP continues until all the objects belong to clusters. Hence, the distribution over clusters for object *i* conditioned on the cluster assignments of objects 1, …, *i*-1 is defined as [Kemp, 2006]:

$$P(z_i = a|z_1, \dots, z_{i-1}) = \begin{cases} \gamma/(i-1+\gamma) & a \text{ is a new cluster} \\ n_a/(i-1+\gamma) & n_a > 0 \end{cases} \quad (3)$$

[Kemp, 2006] explains that *the distribution on z induced by the CRP is exchangeable: the order in which objects are assigned to clusters can be permuted without changing the probability of the resulting partition. P(z) can therefore be computed by choosing an arbitrary ordering and multiplying conditional probabilities. Since new objects can always be assigned to new clusters, the IRM effectively has access to a countably infinite collection of clusters.*

From the clusters generated by the CRP, relations are generated based on the following generative model:

- As described above, for the cluster assignment of objects
  $z \mid \gamma \sim$ CRP (γ)
- For link probabilities between clusters
  $\eta\ (a, b) \mid \beta \sim$ Beta (β, β)
- For links between objects
  $R\ (i, j) \mid z, \eta \sim$ Bernoulli ($\eta\ (z_i, z_j)$)

In the above generative model, we set parameters β=1, and γ=log (*Ji*) where *Ji* is the number of concepts in each mode.

In here, relationships are assumed to be conditionally independent given cluster assignments [Kemp, 2006]. The eventual purpose of the generative model is to identify a cluster *z* that maximizes $P(z|R)$. Based on the generative model defined above, relations from clusters are generated by:

$$P(R|z, \eta) = \prod_{ab} (\eta_{ab})^{m^+_{ab}} (1 - \eta_{ab})^{m^-_{ab}} \quad (4)$$

where m^+_{ab} refers to the total number of links between categorical classes *a* and *b*; and m^-_{ab} refers to the total number of non-links between categorical classes *a* and *b*. The conjugate prior η_{ab} is in the aforementioned generative model defined as: $\eta(a, b) \mid \beta \sim$ Beta(β, β). Accordingly, the conjugate prior η_{ab} is integrated out in the following:

$$P(R|z) = \prod_{a,b\in N} \frac{Beta(m^+_{ab}+\beta, m^-_{ab}+\beta)}{Beta\ (\beta,\beta)} \quad (5)$$

From formulae (3) and (5), the IRM identifies *z* that maximizes: $P\ (z|R) \propto P\ (R|z)\ P(z)$. According to [Kemp, 2006], the expected value of η_{ab} given *z* is:

$$\frac{m^+_{ab} + \beta}{m^-_{ab} + m^+_{ab} + 2\beta} \quad (6)$$

The procedure for the inference is further described in [Mørup 2010]. The solutions displayed in the following are based on the sample with highest likelihood.

4. Results

Based on the BMG and the IRM reviewed above, empirical studies are performed based on the experimental strategies described in Section 2.

4.1 Experimental strategy 1

The first experimental strategy is in a way the most natural approach to judge how an ontology is learned from data consisting of CSCs and features that respectively represent specific domain knowledge existing in two cultures. Thus the IRM is directly applied to the CSC-feature matrices, respectively created from the aforementioned English corpora describing the Danish- and the Japanese educational systems. Accordingly, 59 Danish CSCs and 229 features are simultaneously clustered into 5 and 10 categorical classes in Figure 1-a. In the same way, 54 Japanese CSCs and 229 features are respectively clustered into 6 and 11 categorical classes in Figure 1-b. In both Figure 1-a and 1-b, the unsorted graph shows the relations between the CSCs and their features. It means that each dot represents a relation that a CSC possesses a specific feature in the matrix. The upper right graphs in the figures show the graph sorted according to extracted assignments of clusters computed by the IRM algorithm. The bottom-left graphs show the distribution of CSCs over the extracted categorical classes, and the bottom-right graphs show the distribution of features over the extracted clusters. The bottom-center graphs correspond to the graphs sorted according to extracted assignments of clusters, which indicates the density of relationships between a Danish (or a Japanese) categorical class and a feature cluster. In Figure 2, all members (i.e. specific CSCs) for each categorical class are respectively listed for the Danish and Japanese educational domain knowledge.

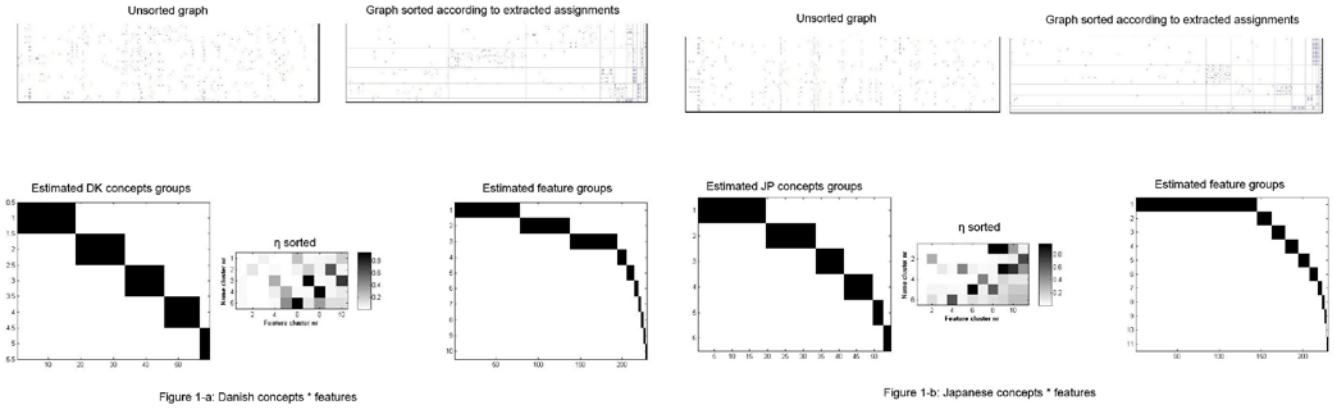

Figure 1: IRM clustering based on the Strategy 1 (Left: Danish CSCs * features / Right: Japanese CSCs * features)

Cluster 1 (Pre primary + adult education)
'D13' continuing education and training for adults
'D14' continuing vocational education and training
'D15' day-care facilities
'D22' further education for adults
'D23' general adult education
'D26' higher preparatory single subjects courses
'D28' higher education for adults
'D32' integrated institution
'D33' Kindergarten (børnehave)
'D38' Nursery/Créche/vuggestue
'D39' open education
'D41' preparatory adult education programme for adults
'D42' pre-primary education
'D43' preschool class
'D49' technical college
'D5' basic adult education
'D55' Vocational college
'D8' business college

Cluster 2 (Primary + alternative primary + Doctoral)
'D1' 10th form
'D12' Continuation school (efterskole)
'D17' doctoral degree
'D18' Dual system
'D19' first stage
'D20' Folk High School
'D21' Full time system
'D30' home tuition
'D36' municipal school/Folkeskole
'D37' non-formal adult education
'D44' private school
'D46' school-leaving examinations
'D47' second stage
'D48' primary and lower secondary education
'D59' youth school

Cluster 3 (Upper secondary)
'D24' general upper secondary education
'D25' gymnasium/STX
'D27' HHX (højere handels eksamen)
'D29' higher preparatory exam
'D31' HTX (højere teknisk eksamen)
'D34' main course
'D54' upper secondary education
'D56' vocational education and training
'D57' vocational upper secondary education
'D58' youth education programmes
'D6' basic course
'D7' Basic vocational education and training

Cluster 4 (Tertiary)
'D10' Centres for Higher Education
'D11' college sector
'D2' Academies of professional higher education
'D3' Academy profession degree programmes
'D40' PhD programme
'D45' professional bachelor programme
'D50' tertiary education
'D51' University Bachelor degree programme
'D52' University college
'D53' University education
'D9' Candidates degree programme

Cluster 5 (open education)
'D16' Diploma programmes
'D35' master programme
'D4' Advanced adult program

Cluster 1 (Upper secondary)
'J1' Advanced course (Senkoka)
'J13' Fisheries course
'J14' Full time course
'J16' General course (Honka)
'J17' Graded course
'J19' Home-economics course
'J2' Agriculture course
'J31' Nursing course
'J32' Ordinary education course
'J33' Part time course
'J44' Public upper secondary school
'J47' Specialized course (Bekka)
'J48' Specialized education course
'J50' Technology course
'J54' Upper secondary school
'J6' Commerce course
'J7' Comprehensive education course
'J8' Correspondence course
'J9' Credit course

Cluster 2 (Tertiary + Primary)
'J11' Doctoral degree
'J12' Elementary school
'J18' Graduate school
'J20' Junior college
'J26' Master's degree
'J28' National elementary school
'J29' National university
'J36' Private elementary school
'J38' Private university
'J39' Professional graduate school
'J40' Public elementary school
'J43' Public university
'J51' Undergraduate department
'J52' University

Cluster 3 (alternative post compulsory)
'J15' General course
'J27' Miscellaneous schools
'J3' colleges of technology
'J4' colleges of technology economy, IT management course
'J46' Specialized course
'J49' Specialized training college
'J5' colleges of technology – industrial course
'J53' Upper secondary course

Cluster 4 (Pre Primary + post compulsory)
'J10' Day care center
'J21' kindergarten
'J22' Kindergarten 1 year course
'J23' Kindergarten 2 years course
'J24' Kindergarten 3 years course
'J30' nursery school
'J34' post-compulsory educational institution
'J35' Pre-school education

Cluster 5 (Secondary)
'J37' Private six-year secondary school
'J42' Public six-year secondary school
'J45' six-year secondary school/ comprehensive secondary school

Cluster 6 (Lower secondary)
'J25' lower secondary school
'J41' Public lower secondary school

Figure 2: CSC members that constitute each categorical class based on Strategy 1 (Left: Danish / Right: Japanese)

For convenience, each categorical class has been named in Figure 2 within a parenthesis based on members that constitute the specific categorical class in question. As Figure 2 shows, some categorical classes (e.g. Danish classes 3, 4 and 5; and Japanese classes 1, 3, 5, and 6) are successfully formed only with CSCs that are related to the respective categorical classes such as "upper secondary", "open education", "secondary", and "lower secondary". However, the rest of the categorical classes are

partly formed with CSCs that represent different categorical classes. For example, the Danish categorical class 1 consists of CSCs that are supposed to belong to "pre primary" and "adult education" and Japanese categorical class 2 consists of CSCs that are supposed to belong to "tertiary" and "primary". When observing Figure 1-a, the successful Danish categorical class 3 "upper secondary" has a very dense relationship with feature cluster 7 consisting of "16-18 years old" and "post compulsory education" and with feature cluster 10 consisting of "upper secondary education" and "vocational perspectives". In the same way; the Danish categorical class 5 representing degree programs targeted for adults has a dense relationship with feature cluster 6 consisting of features: "opportunities for lifelong learning", "part time", "possibilities for combining education and work", "occupational function", and "open education". Figure 1-b shows another notable point that the Japanese categorical classes 1 and 3 both have a dense relationship with feature cluster 9 consisting of "non-compulsory educational school" and "post-compulsory education". However, the Japanese categorical class 1 - "upper secondary" - has also a strong relationship with feature cluster 8 consisting of "16-18 years old". Also the Japanese categorical class 3 - "alternative post compulsory" - has another relationship with feature cluster 10 consisting of "education + practical training". This indirectly indicates that the Japanese categorical classes 1 and 3 both belong to a super-ordinate category (although it does not exist in the dataset) referring to "post-compulsory education". This kind of information could be useful for representing knowledge in a taxonomical structure, e.g. for constructing Terminological Ontologies [Madsen 2004].

The results of experimental strategy 1 indicate that if few decisive features exist for representing a categorical class, the IRM effectively sorts CSCs that relate to these decisive features. However, when relationships between categorical classes and feature clusters are weak, there is a tendency that CSCs that belongs to different categorical classes start to be mixed into one class.

4.2 Experimental strategy 2

The second strategy is to apply the IRM directly to the matrix where the two CSC-feature matrices respectively representing the Danish- and Japanese educational domain knowledge are merged. The aim of this second strategy is to assess whether the IRM can directly be used for mapping CSCs existing in two cultures. Accordingly, in total 113 CSCs (59 Danish and 54 Japanese) and 229 features are respectively clustered into 8 categorical classes and 16 feature clusters as shown in Figure 3. Figure 4 lists 8 categorical classes and their cluster members.

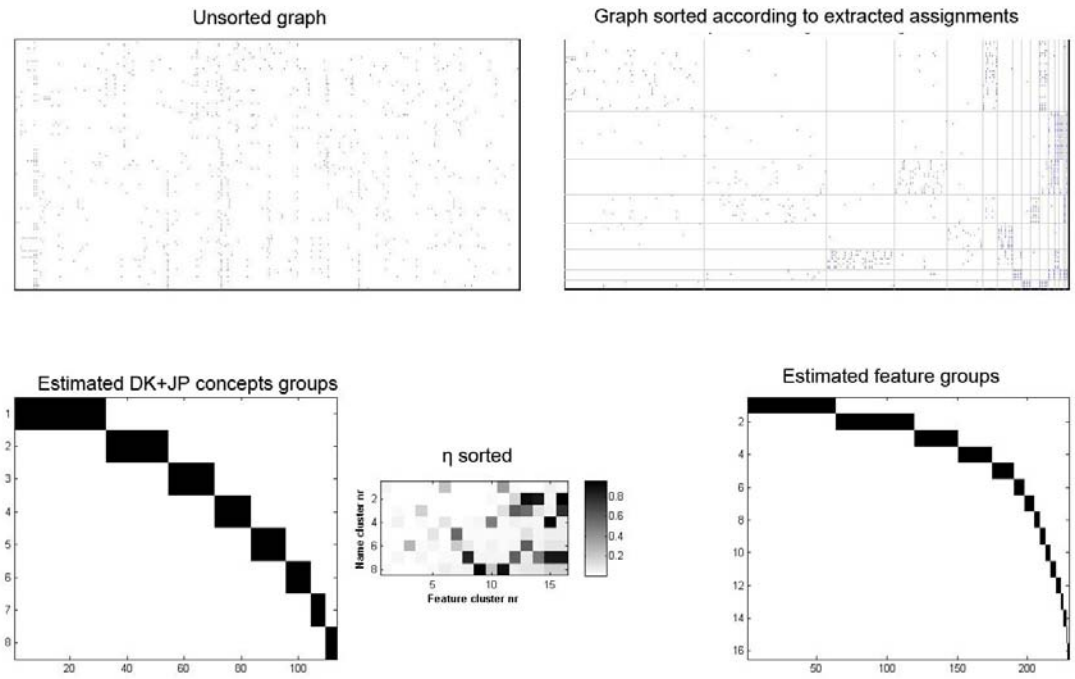

Figure 3: IRM clustering based on Strategy 2 (Danish + Japanese CSCs * features)

The results in Figure 4 show that 3 out of 8 categorical classes are mono-cultural categories. More specifically, class 4 solely consists of Danish CSCs referring to the Danish adult education; and the classes 7 and 8 solely consist of Japanese CSCs respectively referring to the Japanese alternative post-compulsory education and graduate school. It means that the purpose of mapping CSCs existing in the two cultures has not been achieved in terms of these three categorical classes. On the other hand, one interesting finding is that the Danish mono-cultural categorical class 4 and the Japanese mono-cultural categorical class 7 share feature cluster 15 consisting of "opportunities for lifelong learning" according to the η-sorted graph in Figure 3. Although concepts existing in different cultures have not been grouped in the same class, the IRM enables us to identify a relationship indicating which categorical class share which feature cluster in the η-sorted graph. This type

of information is highly useful for constructing feature-based ontologies as well as for CSC-mappings. Finally, the rest of the five bi-cultural categorical classes result in rather ambiguous categories. For example, categorical class 2 that is the group of upper secondary CSCs, consists of the most abstract Danish upper secondary CSCs and all concrete Japanese upper secondary CSCs. On the contrary, categorical class 3 that is also the group of upper secondary CSCs representing concrete Danish upper secondary education for both general and vocational purposes are grouped together with the Japanese college of technologies (JP: *Koto Senmon Gakko*) which provides 5 years of practical and vocational education consisting of 3 years upper secondary and 2 years post-secondary education. Accordingly, the mapping results obtained from the second strategy seems not to be optimal.

Cluster 1 (pre primary + tertiary)
'D10' Centres for Higher Education
'D11' college sector
'D13' continuing education and training for adults
'D15' day-care facilities
'D17' Doctoral degree
'D2' Academies of professional higher education
'D3' Academy profession degree programmes
'D32' integrated institution
'D33' Kindergarten (børnehave)
'D38' Nursery/Créche/vuggestue
'D40' PhD programme
'D42' pre-primary education
'D43' preschool class/0th form/børnehave klasse
'D45' professional bachelor programme
'D50' tertiary education
'D51' University Bachelor degree programme
'D52' University college
'D53' University education
'D9' Candidates degree programme
'J10' Day care center
'J20' Junior college
'J21' kindergarten
'J22' Kindergarten 1 year course
'J23' Kindergarten 2 years course
'J24' Kindergarten 3 years course
'J29' National university
'J30' nursery school
'J35' Pre-school education
'J38' Private university
'J43' Public university
'J51' Undergraduate department
'J52' University

Cluster 4 (adult education)
'D14' continuing vocational education and training
'D16' Diploma programmes/Diplomuddannelse
'D20' Folk High School
'D22' further education for adults
'D23' general adult education
'D26' higher preparatory single subjects courses
'D28' higher education for adults
'D35' master programme
'D37' non-formal adult education
'D39' open education
'D4' Advanced adult program
'D41' preparatory adult education programme for adults
'D5' basic adult education

Cluster 2 (upper secondary)
'D54' upper secondary education
'D58' youth education programmes
'J1' Advanced course (Senkoka)
'J13' Fisheries course
'J14' Full time course
'J16' General course (Honka)
'J17' Graded course
'J19' Home-economics course
'J2' Agriculture course
'J31' Nursing course
'J32' Ordinary education course
'J33' Part time course
'J34' post-compulsory educational institution
'J44' Public upper secondary school
'J47' Specialized course (Bekka)
'J48' Specialized education course
'J50' Technology course
'J54' Upper secondary school
'J6' Commerce course
'J7' Comprehensive education course
'J8' Correspondence course
'J9' Credit course

Cluster 3 (general- and vocational upper secondary + college of technology)
'D24' general upper secondary education
'D25' gymnasium/STX
'D27' HHX (højere handels eksamen)
'D29' higher preparatory exam
'D31' HTX (højere teknisk eksamen)
'D34' main course
'D49' technical college
'D55' Vocational college
'D56' vocational education and training
'D57' vocational upper secondary education
'D6' basic course
'D7' Basic vocational education and training
'D8' business college
'J3' colleges of technology
'J4' colleges of technology – economy, IT management course
'J5' colleges of technology – industrial course

Cluster 5 (primary + lower secondary)
'D1' 10th form/bridge building
'D19' first stage
'D30' home tuition
'D36' municipal school/Folkeskole
'D44' private school
'D46' school-leaving examinations
'D47' second stage
'D48' primary and lower secondary education
J12 elementary school
'J28' National elementary school
'J36' Private elementary school
'J40' Public elementary school

Cluster 6 (alternative lower secondary + secondary school)
'D12' Continuation school (efterskole)
'D18' Dual system
'D21' Full time system
'D59' youth school
'J25' lower secondary school
'J37' Private six-year secondary school
'J41' Public lower secondary school
'J42' Public six-year secondary school
'J45' six-year secondary school/ comprehensive secondary school

Cluster 7 (alternative post compulsory)
'J15' General course
'J27' Miscellaneous schools
'J46' Specialized course
'J49' Specialized training college
'J53' Upper secondary course

Cluster 8 (graduate school)
'J11' Doctoral degree
'J18' Graduate school
'J26' Master's degree
'J39' Professional graduate school

Figure 4: CSC members that constitute each categorical class based on Strategy 2 (Danish + Japanese concepts)

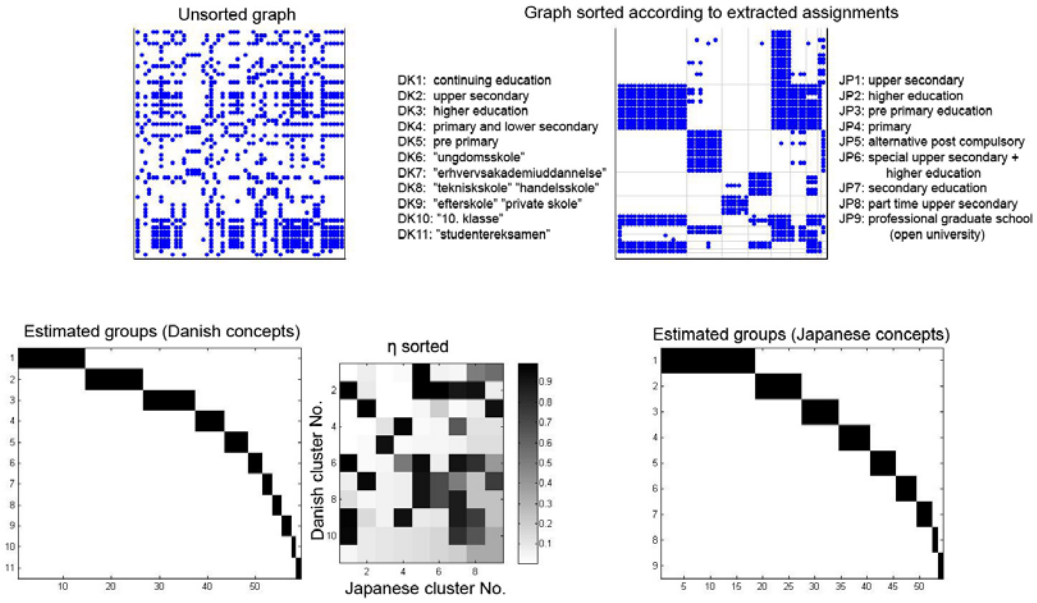

Figure 5: IRM clustering based on the Strategy 3 (BGM + IRM)

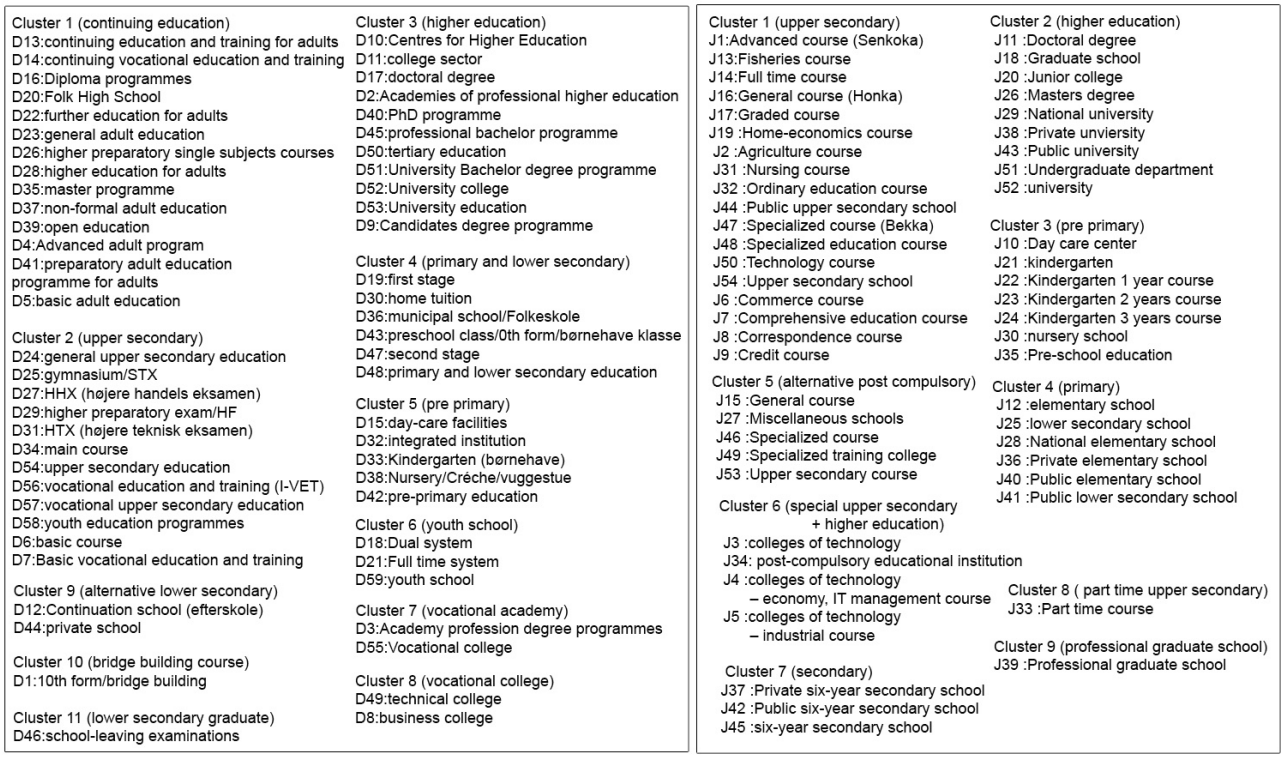

Figure 6: CSC members that constitute each categorical class based on Strategy 3 (Left: Danish / Right: Japanese)

4.3 Experimental strategy 3

The third strategy is to apply the BMG to directly compute similarity relations between CSCs existing in the two cultures, and thereafter to apply the IRM in order to cluster CSCs in the respective countries into categorical classes. This enables us not only to observe the inter-relations of categorical classes existing in the two cultures, but also to instantly scrutinize more specific similarity relations between each category member (i.e. CSCs) existing in the two cultures. Accordingly, 59 Danish CSCs and 54 Japanese CSCs are simultaneously clustered into 11 and 9 categorical classes, respectively shown in Figure 5. In this figure, the unsorted graph shows existing links between the Danish CSCs and the Japanese CSCs identified by the computation of the BMG. It means that each dot represents a link established between a Danish CSC and a Japanese CSC, when they share at least one common feature. The upper right graph in Figure 5 shows the graph sorted according to extracted assignments of categorical classes computed by the IRM algorithm. The bottom left- and right graphs show the distribution of concepts over the extracted categorical classes, respectively for the Danish- and Japanese CSCs. The bottom center graph corresponds to the graph sorted according to extracted assignments of categorical classes, which indicates the density of relationships between a Danish categorical class and a Japanese categorical class.

The results in Figure 6 show that both the Danish- and the Japanese CSCs are clustered into a more fine-grained level compared with the results obtained from the first- and second experimental strategies. Almost all members in each categorical class in Figure 6 are grouped together with other members that are supposed to belong to the same categorical class. For example, some CSCs such as the Japanese "J3: college of technology (JP: *koto-senmon-gakko*)" and the Danish "D36: municipal school (DK: *folkeskole*)" are CSCs that are difficult to be categorized in a multi-cultural context. While, in the first- and second experimental strategies, these CSCs have been included in a more ambiguous larger categorical class where CSCs that are supposed to belong to different categorical classes have been grouped together, J3 and D36 are respectively grouped into a more specific and independent categorical class, i.e. the Japanese categorical class 6 and the Danish categorical class 4, in this third strategy. One of the noteworthy points in the third strategy is that, when observing the η-sorted graph in Figure 5, it is possible to study more complex relationships of categorical classes in a cross-cultural context. For instance, the Japanese categorical class 6 where "J3: college of technology" belongs, has a strong relationship with the Danish categorical class 2 "upper secondary" class, but also has a little weaker relationship with both the Danish categorical classes 7 and 8, which respectively represent "vocational academy" and "vocational college" categories providing a 2 years post-secondary degree in Denmark. The observation of the η-sorted graph in Figure 5 further provide a clear picture of how each country-specific categorical class is related to categorical classes existing in another country in a very complex and comprehensible manner. This kind of overview of how categorical classes in different cultures are inter-related is highly useful and valuable not only for mapping CSCs but also for constructing ontologies in a multi-cultural context.

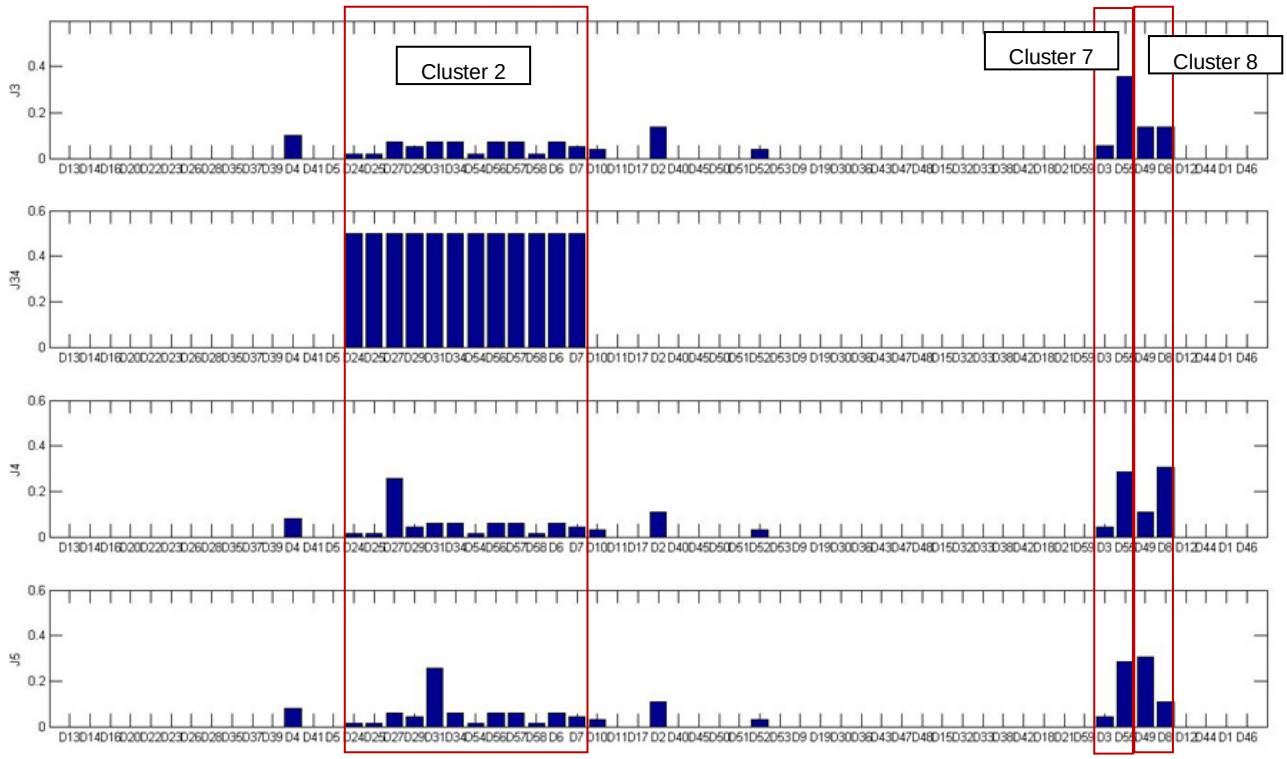

Figure 7: BMG similarity relations between the category members of Japanese categorical class 6 and all the Danish concepts

5. Discussions and future perspectives

The experimental results presented in this work indicate that the third strategy seems to be the most effective approach for clustering CSCs into more specific and appropriate categorical classes. In addition, this strategy enables us to capture complex relationships existing between each categorical class in the two cultures. However, the drawback of the third strategy is that it is not able to assess how each categorical class is related with features. On the other hand, the first- and second strategies enable us to analyze how features and each categorical class are related to each other. The shortcoming of these strategies is that the clustered categorical classes are rather ambiguous and some categorical classes are mixed with members that are supposed to belong to other categorical classes.

Another advantage of the third strategy is that the direct application of the BMG enables us to analyze further specific similarity relations between category members of the respective categorical classes existing in the two cultures. Figure 7 illustrates how the category members of the Japanese categorical class 6 in Figure 6 are related with the category members of the Danish categorical classes 2, 7 and 8. As discussed in the previous section, the η-sorted graph in Figure 5 shows that the Japanese categorical class 6 has the strongest relationship with the Danish categorical class 2 and slightly weaker relationship with the Danish categorical classes 7 and 8. Figure 7 explains these relationships between the classes by showing that all the category members of the Danish categorical class 2 share at least one feature with all the category members of the Japanese categorical class 6, while only 75% of the category members of the Danish categorical classes 7 and 8, respectively, share at least one feature with 75% of the category members of the Japanese categorical class 6. On the other hand, when observing individual relationships between category members between the Japanese- and the Danish categorical classes, similarity relationships are not necessarily strong in most of the combinations. Here, the selection of feature-based similarity measures plays in to the considerations.

In this work, we have selected the BMG as the most suitable feature-based similarity measure. However, for implementing the IRM based on the third experimental strategy, it is possible to apply other feature-based similarity measures, such as the Jaccard similarity coefficient [Tan, 2005], [Jaccard, 1901] and Tversky´s set-theoretic model [Tversky, 1977], which compute similarities based on common features and distinctive features possessed by the two CSCs in question. Comparative qualitative analyses of applying different feature-based similarities to CSC-mapping are further discussed in [Glückstad, 2012-a] and [Glückstad, 2012-b], and our arguments of applying the BMG from both cognitive- and pragmatic point of views are discussed in details in another session of this conference [Glückstad, 2012-c]. Thus, in this work, we focus on how to interpret the results obtained from the BMG shown in Figure 7. As explained in Section 2, equation (1) computes the conditional probability that a new observed object y falls under a consequential region C given the learner's prior knowledge that have already been observed as x. It means that in relation to Figure 7, a scenario could be that a Japanese person who has prior knowledge of the Japanese educational system is learning the Danish educational system by comparing similarities of individual Danish educational CSCs. Thus, all the knowledge about Japanese CSCs are considered as x and the individual Danish CSC as y in

equations (1) and (2) in Section 2. In other words, it can be interpreted that all definitional features possessed by a Japanese CSC are considered as prior knowledge of the Japanese learner and can act as noise (or cultural bias) if a feature is not possessed by a Danish concept taken in comparison. In addition, the uniqueness of the BMG is to reflect the importance of features for the similarity computation by considering features that are possessed by many concepts as less important and features that are possessed by fewer concepts as more important and decisive. In case of "J5: college of technology – industrial course (JP: *Koto-senmon-gakko, sangyo-koosu*)", Danish CSCs - "D31: HTX (Danish upper education that is specialized in technical and natural science)", "D55: vocational college (Danish educational institution that offer vocationally-oriented upper- and 2 year post-secondary education)", and "D8: technical college (Danish vocational college specialized in technical and natural science)" – they are identified as the most similar concepts. In the respective cases, one or few decisive features such as "specialized in technical and natural science", "offering 2 year post-secondary degree" strongly influence the similarity computation and differentiate the confidence levels of similarity from other concepts. Another interesting pattern of applying the BMG as shown in Figure 7 is related to "J34: post-compulsory educational institution (JP: *Gimu-kyoiku-go-no-kyoiku-kikan*)". The J34-CSC is a very abstract concept only possessing two features. It means that no other distinctive features influence as noise in the similarity computation. Accordingly, the result indicates that the Japanese CSC, J34, likely covers all category members of the Danish categorical class 2 as the most similar concepts identified in the Danish culture. In this way, the combination of the BMG and the IRM could possibly be used for first capturing the abstract relationship between categorical classes and next for further analyze the individual similarity relationship between CSCs in order to achieve more fine-grained CSC-mappings.

A critical point in this work is that the datasets have been created in a way where the authors systematically but manually extracted CSCs and definitional features from the applied text corpora. Although the procedure has been systematized, it is hence not possible to perfectly eliminate human subjectivities. Accordingly, one of our future challenges is to further investigate what types of datasets are suitable for CSC-mapping applying the BMG+IRM solution. One possibility would be to apply this solution to a more complex dataset obtained from other major multilingual ontology projects such as the Monnet project [Declerck, 2010]. Another possibility is to apply the BMG+IRM solution to datasets consisting of CSCs and features that are automatically extracted from text corpora [Lassen, 2012]. This may eventually lead to not only an automated CSC-mapping but also to an automatic ontology learning applying the IRM.

Another aim of this work is, as briefly mentioned in the previous paragraph, to investigate possibilities of applying the IRM for efficiently constructing feature-based ontologies that are multi-culturally interoperable. From this viewpoint, the method of constructing Terminological Ontologies (TOs) in [Madsen, 2004] that are in accordance with [ISO 2000] could be highly suitable for the IRM application. The uniqueness of the TO method is its feature specifications and subdivision criteria [Madsen 2004], [Madsen 2005]. While the use of feature specifications is subject to principles and constraints, the TO-approach allows for so-called poly-hierarchy structures. It means that one CSC may be related to two or more super-ordinate CSCs. Accordingly, the IRM could potentially be used as an effective pre-processing step of constructing TOs that are interoperable in a multi-cultural context. Mainly for two reasons: **1)** the IRM may indicate which features influence the formation of categorical classes as the results from the first- and second experimental strategies in this work have shown; and **2)** the IRM based on the third experimental strategy may cluster CSCs into more specific and appropriate categorical classes that may capture complex relationships between each categorical classes existing in the two cultures. Hence, we need a solution to achieve these strategies at one time. Here, it is important to remind that the design we have chosen for this work is the simplest design of the IRM dealing with two types with a single two-place relation R: $T_1 \times T_2 \rightarrow \{0, 1\}$, and as described in [Kemp, 2006], the IRM design can be a more complex model clustering three relations simultaneously. Accordingly, our obvious future challenge would be to investigate what kind of complex model is applicable for constructing ontologies in collaboration with the automatic TO construction project [Madsen, 2010].

6. Conclusions

In this work, we investigated the application of the IRM [Kemp 2006] to the loosely-structured datasets consisting of CSCs and features representing two cultures for the purpose of mapping CSCs in a multi-cultural context. The results from the three applied experimental strategies indicate that the combination of the BMG and the IRM seems to be the most effective approach for not only clustering CSCs into more specific and appropriate categorical classes but also for capturing complex relationships between each categorical classes existing in the two cultures. In addition, the direct application of the BMG to the datasets enables us to effectively analyze further specific similarity relations between category members existing in the two cultures. However, in order to conclude on this, it is necessary to investigate the performance of the BMG+IRM solution with different types of datasets that are purely objectively generated. Another important finding is the potential application of the IRM for the automatic construction of feature-based ontologies among others, TOs [Madsen 2004]. The results obtained from the first- and second experimental strategies indicate that this may potentially be achieved by designing a more complex IRM. Although further research is required, the application of the IRM to the multi-cultural ontologies seems to provide a diverse potential in this research domain.

Acknowledgement

We would like to express our thanks to the DanTermBank project members, among others, Bodil Nistrup Madsen, Hanne Erdman Thomsen and Tine Lassen at the Copenhagen Business School for valuable and relevant discussions related to our work.

References

[Chomsky 1986] Chomsky, N., Language and Problems of Knowledge: The Managua lectures. Cambridge, MA: MIT Press, 1986

[Cimiano 2010] Cimiano P., Montiel-Ponsoda E., Buitelaar P., Espinoza M., Gómez-Pérez A. A Note on Ontology Localization., Journal of Applied Ontology Vol. 5, No. 2, IOS Press, 2010.

[Declerck 2010] Declerck T., Krieger H.U., Thomas S.M., Buitelaar P., O'Riain S., Wunner T., Maguet G., McCrae J., Spohr D., Montiel-Ponsoda.E. Ontology-based Multilingual Access to Financial Reports for Sharing Business Knowledge Across Europe., In Internal Financial Control Assessment Applying Multilingual Ontology Framework, J. Rooz, J. Ivanyos, Eds. Budapest: HVG Press, 2010

[de Souza 2004] de Souza K.X.S., Davis J., Aligning Ontologies and Evaluating Concept Similarities, In On the Move to Meaningful Internet Systems 2004: Lecture Notes in Computer Science, Volume 3291, Springer, 2004.

[Espinoza 2009] Espinoza, M., Montiel-Ponsoda, E., Gómez-Pérez, A. Ontology Localization. In Proc. the 5th International Conference on Knowledge Capture (KCAP) New York, NJ, USA, 2009

[Glückstad 2012-a] Glückstad F.K., Mørup M., Feature-based Ontology Mapping from an Information Receivers' Viewpoint, (under review)

[Glückstad 2012-b] Glückstad F.K., Cross-Cultural Concept Mappings of Standardized Datasets, (under review)

[Glückstad 2012-c] Glückstad F.K., Bridging Remote Cultures: Influence of Cultural Prior-Knowledge in Cross-Cultural Communication, In Proc. The 26th Annual Conference of the Japanese Society for Artificial Intelligence, Japan, June 2012.

[Hansen 2011] Hansen T.J., Mørup M., Hansen L. K., Non-parametric Co-clustering of Large Scale Sparse Bipartite Networks on the GPU, In IEEE International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2011.

[Huang 2010] Huang H.H., Kuo Y.H., Cross-lingual Document Representation and Semantic Similarity Measure: A fuzzy set and rough set approach, In IEEE Transaction on Fuzzy Systems, Volume 18:6, IEEE, 2010.

[ISO 2000] ISO 704. Terminology Work — principles and methods, ISO, 2000

[Jaccard 1901] Jaccard, P., Étude comparative de la eistribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles 37: 1901

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

[Lassen 2012] Lassen, T., A Corpus Compilation and Processing Prototype for Terminology Work., In Proc. the 10th International Conference on Terminology and Knowledge Engineering, Spain, June 2012.

[Madsen 2004] Madsen, B.N., Thomsen, H.E. and Vikner, C., Principles of a System for Terminological Concept Modelling. In Proc. the 4th International Conference on Language Resources and Evaluation. ELRA, 2004.

[Madsen 2005] Madsen, B.N., Thomsen, H.E. and Vikner, C., Multidimensionality in Terminological Concept Modelling. In Proc. of the 7th International Conference on Terminology and Knowledge Engineering, Copenhagen, 2005

[Madsen 2010] Madsen N.B., Thomsen H.T., Halskov J., Lassen T., Automatic Ontology Construction for a National Term Bank, In Proc. Terminology and Knowledge Engineering Conference 2010, Dublin, 2010.

[Mørup 2010] Mørup M., Madsen K.H., Dogonowski A.M., Siebner H., Hansen L.K., Infinite Relational Modeling of Functional Connectivity in Resting State fMRI., In Neural Information Processing Systems, 2010.

[Pitman, 2002] Pitman J., Combinatorial Stochastic Processes. Notes for Saint Flour Summer School, 2002.

[Shepard 1987] Shepard R.N., Towards a Universal Law of Generalization for Psychological Science. Science, 237, 1987.

[Tan 2005] Tan P.N., Steinbach M., Kumar V., Introduction to Data Mining, Pearson Education, Inc. 2006.

[Tenenbaum 2001] Tenenbaum, J. B., & Griffiths, T. L., Generalization, Similarity, and Bayesian Inference. Behavioral and Brain Sciences, Vol.24(4; 4), 2001

[Tversky 1977] Tversky, A., Features of Similarity. Psychological Review, Vol., 84(4; 4), 1977

[Vossen 2008] Vossen P., Agirre E., Calzolari N., Fellbaum C., Hsieh S., Huang C.R., Isahara H., Kanzaki K., Marchetti A., Monachini M., Neri F., Raffaelli R., Rigau G., Tescon M., VanGent J. KYOTO: A system for mining, structuring and distributing knowledge across languages and cultures. In Proc. the 6th International Conference on Language Resources and Evaluation, Morocco, 2008.

[Xu 2006] Xu Z., Tresp V., Yu K., Kriegel H.P., Infinite Hidden Relational Models, In Proc. 22nd Conf. on Uncertainty in Artificial Intelligence (UAI'06) Cambridge, MA, 2006