

A Simple Metric to Measure Semantic Overlap between Models: Application and Visualization

Jean-Paul Van Belle

Department of Information Systems, University of Cape Town, South Africa, jvbelle@commerce.uct.ac.za

Abstract.

This paper investigates a fairly simple but easily automatable metric for measuring the semantic overlap between models as a proxy to the degree of overlap between their domain coverage. Such a metric is very useful when evaluating competing models with fully or partially overlapping domains, be it for purposes of model integration, re-use or selection. The proposed metric is based on the semantic information contained within the models and inspired by computational linguistics and ontology research. Because pair-wise comparison of multiple models results in large tables, some visual techniques are suggested to facilitate the analysis process. The proposed metric and visualizations are illustrated and validated by applying them to a set of real-world enterprise data models.

1. Introduction

The enterprise domain can be modelled in many alternative ways. The exact contents and appearance of an enterprise model will depend on a wide range of contextual variables including the methodology and modelling notation used, the reference discipline, the purpose of the model, and the cognitive analysis process used by the modeller(s). Because the IT industry is moving increasingly towards (re-)using commercially available pre-developed models, alternative models are available which cover fully or partially the same domain area. Modellers thus need to compare the degree of overlap between these models. Alternatively it may be useful to compare in-house models against each other (or outside models). This is especially relevant when two merging enterprises or business units wish to integrate their information systems or enterprise information architectures. Finally, many upper-CASE tools developers may wish to add functionality that allows system modellers to compare models with each other, for instance for version control or model integration.

This paper aims to present a pragmatic approach towards measuring the degree of similarity or overlap between models, by using a fairly simple metric which uses only the semantic information contained within the models. Although measuring semantic similarity is a huge and complex research area, the pragmatic overall purpose of the research was to find and validate a metric which, while by no means

perfect, will be good enough to work in the real world and can be automated easily.

The key objectives of the paper are thus (1) to develop a metric for measuring the semantic overlap between models which can easily be fully automated; (2) to provide some means of visualizing and interpreting the metric; and (3) to test and illustrate the proposed metrics and analysis techniques using a sample of medium-sized (static) enterprise models.

2. Prior Research

Systems engineering researchers have devised many metrics for model analysis based on syntactic or structural model information: de Marco's Bang, McCabe's cyclomatic complexity, connectivity, reuse ratio, inheritance depth, graph size, etc. However, very little work has been done on the *semantic* analysis of models. An interesting, structural approach towards model comparison and integration was presented by Chen-Burger [1;2] where a "heuristic similarity assessment function" was used to "quantify the quality of a match" between entities in their "Generic Model Advisor". Their function includes the matched structural relationships instead of attributes but requires significant human input to identify matches. By contrast, our approach suggested below can be fully automated (no user interaction is required) and has a more sophisticated approach in terms of mapping similarity using meaning rather than structural similarity.

On the other hand, a large amount of research has been done in the area of ontology research concerning the integration or merging of ontologies [3;4]. Other relevant research has happened in the field of categorization theory [5] and information theory [6]. Similarly, in linguistic text analysis, much research exists on measuring document similarity [7; 8]. This research has come to the foreground in the context of search engine algorithms [9] and document translation [10]. The techniques presented below borrow from the research in ontology merging and document similarity.

This paper actually forms part of a larger research effort which looks at a comprehensive framework for analysing models [11]. This framework proposes a

number of other semantically-based metrics: the use of meta-models, expert knowledge and reference frameworks for comparing the coverage and overlap between models. However, these other approaches required significant human input, were very subjective and their overall validity remained relatively suspect.

3. Methodology

In principle, the approach to measuring the degree of model similarity or overlap is simple. What is required is to (attempt to) map each model element of a particular model to a semantically equivalent model element in the other model(s). Where corresponding model elements are found, these increase the similarity between the models; the more model elements for which no corresponding element can be found in another model, the greater the difference (semantic distance) between the models concerned.

The overall methodology can be broken down into three distinct steps, each posing unique challenges. Each step corresponds to one section of the paper.

1. Where possible, map/link the model elements from the (source or base) model to the corresponding entities in (target) model(s). This step involves semantic processing to account for the fact that different word labels (or tokens) can refer to the same underlying concept. This is discussed in section 5.
2. Based on the mappings, calculate a similarity (or distance) metric between the various models. This involves choosing from a number of distance measures and performing a large number of calculations, typically resulting in a matrix-like table with distance values (Section 6).
3. Analyse and interpret the similarity indices. Because the distance tables represent a multi-dimensional space (e.g. 22 dimensions for 23 models), they are difficult to interpret at first sight. Further statistical transformations are required to visualize the findings (Section 7).

A somewhat similar approach was suggested in [12] who investigated the similarity between documents, but a more sophisticated approach will be presented below, in order to deal with the problem of synonymy. Although the correspondence analysis should be applied to all model elements (including relationships and groupers), this trivial extension was not illustrated below due to the fact that most models in the database did not name their relationships or grouper constructs. Accounting for model structure in terms of super/subclass hierarchy is much more

complex and can only be accounted for in a limited way by using hypernymy.

4. The Database of Models

In order to demonstrate the practical value of the metrics for real-world, industrial-strength systems, twenty *enterprise data models* were captured, more than half of which have a *concept count* exceeding 1000. Extending the proposed analysis techniques to *process* models is not necessarily trivial, but the underlying principles remain the same.

For validity purposes, the models in the database are grounded in a wide variety of reference disciplines and methodologies. This demonstrates the feasibility of the metric and its analysis across modelling notations. Some of the models are in wide-spread use, such as the SAP and BAAN models which serve as the basis for internationally adopted Enterprise Resource Planning (ERP) systems, and the data model libraries which are well-known in the system engineering practitioners' community. Some lesser known models were also included with the specific purpose of increasing the variety and quality range within the model database. The database with the captured models in XML format is available from the author to other researchers on simple request. The following models were included.

- Two ERP models: the SAP R/3 [13] and BAAN IV [14].
- Four published enterprise data model libraries: BOMA [15], Fowler's analysis patterns [16], Hay [17] and Silverston [18].
- Two academic reference models: ARRI's Small Integrated Manufacturing Enterprise Model and Purdue's Reference Model for CIM [19].
- Three enterprise ontologies: the enterprise ontology developed by AIAI [20], TOVE from EIL and a subset of OpenCyc (all organisation and enterprise-related concepts).
- Some lesser know models: Inmon's set of models for data warehousing, the Belgian accounting framework [21], a financial spreadsheet model [22], AKMA's Generic DataFrame, San Francisco (predecessor of IBM's WebSphere), Miller's General Living Systems Model as an exemplar of systems theory [23], the NHS's Generic Health Care Management Class Model as an example of an industry-specific model, Nippon Steel's small DFD model for CIM [19], a randomly generated wordlist and two subsets of the Ottawa Business Journal's hyperlinked business dictionary (no longer available on www.ottawabusinessjournal.com) as examples of semantic models.

A number of methodological issues arose in the capture of the models, including the selection of an appropriate meta-model, harmonizing the definitions of concepts between different modelling paradigms and the very different degrees of formality between models; a full discussion is found in [11]. However, these issues do not impact the findings materially.

5. Mapping the Semantic Correspondence between Model Elements

Mapping correspondence between model elements involves analysing every model entity and checking whether it can be mapped to an equivalent element in another model. The simplest approach is to measure the literal overlap between models. This approach looks for model elements (entities, relationships, groupers, attributes) between models that have the exact same label or name (word token): most enterprise models use common domain names for entities such as “customer”, “payment”, “worker” etc. The degree of direct model overlap between any two models can then be found by counting the exact matches. Because this number depends on the overall model size, it is best expressed as a percentage of the total model size – either of the selected referent model or alternatively as the average size of both models.

This direct overlap was calculated for all possible unique ($23 \times 22 / 2 = 253$) model pairs in the database. Since most models did not include relationship names or attributes, the analysis below was only applied to model entity/object names. As can be expected, the “direct” overlap between models is fairly low with calculated overlap ratios rarely exceeding 10%.

However, this naïve approach has some major methodological issues. Firstly, the same work (token) can indicate a different concept i.e. words carry different meanings. Although this could potentially be picked up by looking at the attributes and relationships of the concept, it is a fairly intractable problem [24]. The focus in the remainder of this paper will be on the converse case – that of synonymy – which is believed to be a much more common problem. Model(ler)s could use synonyms i.e. different word tokens to indicate the same concept. Indeed, [12] reports that the chance of two (experienced) modellers using the same term for a given domain concept is less than 20%.

One possible approach, suggested in [3] for the purpose of merging ontologies, requires significant input by a human expert. However, it remains a

subjective and very expensive process. Another, possibly automatable approach for trying to cope with synonyms can be found in [24] who uses topic signatures based on extensive analysis of web pages returned by search engines. However, this approach is computationally quite expensive for large vocabulary sets. Hence an objective, algorithmic and computationally efficient approach is presented to determine a match between synonyms – concepts that have the same meaning but use different word tokens. It relies on a key resource developed by computational linguists to deal in a methodologically defensible way with synonymy: WordNet [25].

WordNet is “an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets (synsets), each representing one underlying lexical concept. Different relations link the synsets.” [<http://wordnet.princeton.edu>]. This database and its associated software were developed by the Cognitive Science Laboratory at Princeton University and have been made available for linguistic and related research. It is widely used in the linguistics research community. WordNet also includes other lexical relationships, including hypernyms: more general terms equivalent to super-classes. This enables specializations to be mapped against generalizations, even if the modelling language does not allow generalizations, and these concepts were therefore excluded from the model.

For this research, WordNet 1.6 was used to generate synonyms and hypernyms for each word of each model entity. In an effort to exclude the more obscure meanings or word uses, only the synonyms and immediate hypernyms for the *first two most frequently used* noun senses as listed in WordNet were used, and no more than the ten most frequently used synonyms were considered. It is believed that this greatly reduced the amount of noise while excluding less than 4% of the synonyms. The procedure of matching words was repeated but now matching could also be done against synonyms and hypernyms in the target model. No attempt was made to map the synonym for an entity against a synonym of the other model because that could create spurious or even incorrect mappings. Where two different original entities (words) from the one model were mapped to one single entity in the other model, the “model overlap” was calculated as the average of the two values.

Calculating the overall overlap between models based on the synonym-enhanced metric leads to a marked increase in the number of mappings (correspondence) between models. For instance, the 32% of the entities were identical in both Baan and SAP (the models are roughly the same size). When synonyms are taken into account, 47% of the entities can be mapped. For smaller models, the improvement is often even more dramatic. For example, using synonyms increases the model overlap between AKMA and AIAI from 10% to 24% for AKMA and from 9% to 21% for AIAI. Across all models, using synonyms improves the average overlap between models from 14% to 26% - almost a doubling!

Future research could refine the concept of the “binary” matching of synonyms (whether or not the entity from one model is identical to the synonym of an entity of another model) with a “fuzzy” match which can take on a value ranging between a full to no match. One approach could be based on attributes that corresponding entities from different models have in common, as suggested in [26] and [27]. Because very few models in the database included attributes, this approach could not be validated in this research. Where available, the descriptions of the various model elements (entities, relationship etc.) could be taken into account rather than just mapping the element names. However, exploratory analysis showed that there tended to be more convergence between model element names than their descriptions.

6. Calculating the Similarity Metrics

This section suggests how the degree of overlap can be quantified by using a similarity or distance metric. In what follows, the term “*similarity*” will be used as denoting the opposite of “*distance*” i.e. shorthand for the degree of *semantic overlap*. It is acknowledged and should be realised that this is a restrictive interpretation of model similarity (see note above).

A fairly large number of similarity measures exist, e.g. the Statistica software package computes the following similarity measures: matching, Jaccard,

Russell & Rao, Hamman, dice, antiDice, Sneath & Sokal, Rogers & Tanimoto, Ochiai, Yule, Anderberg, Kulczynski, Gower2 and Pearson distances. Many of these have not been validated in a linguistic context and many cannot be used in the case where vectors are of unequal length (the word lists for each model are different sizes). Linguists appear to favour three similarity measures in particular: the cosine, dice, and Jaccard distance [5; 6; 7; 8; 9;28]:

$$\text{Cosine Distance} = 1 - \frac{\text{Size(Overlap)}}{\sqrt{[\text{Size(A)} * \text{Size(B)}]}} \tag{1}$$

$$\text{Dice Distance} = 1 - \frac{2 * \text{Size(Overlap)}}{[\text{Size(A)} + \text{Size(B)}]} \tag{2}$$

$$\text{Jaccard Distance} = 1 - \frac{\text{Size(Overlap)}}{[\text{Size(A)} + \text{Size(B)} - \text{Size(Overlap)}]} \tag{3}$$

There does not appear to be any convincing conceptual argument about which of these three measures gives the best results, except for an empirical study in linguistic analysis which suggested that the Jaccard similarity measure may yield better predictive matching results than the cosine or dice measure [29]. It was found in our research that, although the three metrics yield different absolute values, the overall ranking of these values remained almost identical.

In all cases, one can calculate either a *distance* measure, whereby 1 (or 100%) means a far away as possible and 0 means identity or perfect similarity; or the complementary *similarity* measure, which is equal to 1 minus the distance coefficient i.e. 1 = maximal similarity and 0 = no similarity. In this research, similarity measures are used.

Table 1 gives the dice similarities for the subset of the fourteen largest models. Cosine and dice distances have been omitted due to space limitations. Note that the similarity measures used are the ones suggested by [30] i.e. the formula is adjusted for unequal vector sizes. The coefficients are symmetrical i.e. the similarity from model A to model B is the same as from model B to A.

Table 1: Dice Similarity Coefficients for Selected Models Based on Synonyms

Model	AI	AK	AR	BA	BO	CY	FO	HA	IN	PU	SA	SF	SI
TO: Tove	17%	11%	16%	24%	21%	25%	17%	23%	18%	12%	23%	18%	18%
SI: Silverston	25%	29%	24%	43%	44%	26%	38%	47%	36%	31%	43%	41%	
SF: SanFrancisco	25%	31%	20%	38%	35%	24%	36%	32%	26%	25%	35%		
SA: SAP	27%	26%	33%	48%	45%	34%	33%	47%	36%	36%			
PU: Purdue	17%	17%	31%	32%	30%	19%	25%	36%	28%				
IN: Inmon	16%	22%	20%	39%	33%	33%	23%	37%					
HA: Hay	30%	26%	29%	45%	45%	30%	39%						
FO: Fowler	30%	25%	22%	37%	34%	23%							

CY: Cyc	20%	18%	18%	34%	30%
BO: Boma	27%	26%	29%	43%	
BA: Baan	26%	24%	29%		
AR: ARRI	24%	19%			
AK: AKMA	23%				

When comparing the similarity coefficients, one finds that the Jaccard similarity coefficients yield fairly low numbers, generally about half the value of the cosine and dice coefficients. The cosine and dice coefficients are actually very close to the “relative overlap” percentages. Due to the way the Jaccard and dice coefficients are calculated, they preserve relative ranking almost perfectly; thus ranking tables for Jaccard and dice similarity coefficients are virtually the same, despite the fact that the actual values are quite different.

A problem with merely calculating the similarity indices between models is the large volume of data: a metric is calculated for each of the possible pairs of models, leading to a table with n^2 cells (with n = number of models). In order to make the data more accessible, it is necessary to pursue the similarity analysis by means of further statistical methods.

7. Analyzing the Similarity Coefficients using Ranking and Hierarchical Tree Analysis

The similarity indices lend themselves to a variety of analysis techniques and approaches.

7.1 Most similar models

The first analysis consists of looking at the most similar models. This is done by ranking the similarity coefficients in (the full 23-model version of) Table 1. Note that the degenerate values of perfect similarity for the diagonal cells (each model is 100% similar to itself) should be ignored. The following observations can be made:

- The two most similar models are the SAP and the Baan model. Both are ERP models of roughly the same size and functionality and this is a strongest confirmation of the validity of the technique or methodological procedure that has been adopted.
- Almost equally similar are the Hay and Silverston models. Again, this validates the measure since both models come from the same reference discipline (data model libraries) and were published in the same year.
- From there, there are a number of different combinations between the above, which all seem to form a fairly close or similar cluster of models. Refer to [11] for a further discussion. The smaller models were omitted from tables 1 and 2 for space reasons.

Table 1 can be re-arranged to checking for *each model* what its most similar neighbours are. Table 2 lists the closest three neighbours for a number of models using dice similarity. Some interesting additional observations emerge from the above analysis:

- The closeness between ARRI and Nippon. Both models are inspired by the CIM/Enterprise Engineering reference discipline.
- Similarly, the model closest to the Purdue model is the Nippon model. Although no formal “intellectual credit” is mentioned, the Nippon model appears in the appendix of the Purdue model documentation!
- Fairly dissimilar but still the closest neighbour to the TOVE model is CYC, which also stems from the same reference discipline: ontology research.

Surprisingly, the AIAI model is closer to the data models (Fowler, Hay, BOMA) than any other model. Finally, it is interesting – and heartening – to note that the choice of similarity measure does not really affect the analysis. Although there are some differences in the rankings, all of the above remarks are valid regardless of whether the cosine, the dice or the Jaccard coefficients are used.

Table 2: Three Most Similar Neighbours for Each Model (Dice similarity)

Model	Most similar neighbour		2 nd nearest neighbour		3 rd nearest neighbour	
	1	Sim	2	Sim	3	Sim
AI	FO	30%	HA	30%	BO	27%
AK	SF	31%	SI	29%	HA	26%
AR	NI	35%	SA	33%	PU	31%
BA	SA	48%	HA	45%	SI	43%
BO	HA	45%	SA	45%	SI	44%
CY	BA	34%	SA	34%	IN	33%
FO	HA	39%	SI	38%	BA	37%
HA	SA	47%	SI	47%	BO	45%
IN	OB	42%	BA	39%	OD	39%
NI	SA	41%	PU	38%	HA	36%
PU	NI	38%	HA	36%	SA	36%
SA	BA	48%	HA	47%	BO	45%
SF	SI	41%	BA	38%	FO	36%
SI	HA	47%	BO	44%	SA	43%
TO	CY	25%	BA	24%	HA	23%

7.2 Similarity Dendrogram

The above analysis can be extended and confirmed visually in a more systematically way by employing a more appropriate tool from the cluster analysis arsenal: hierarchal tree construction. Because [29] suggest that the Jaccard similarity coefficient is

empirically the best, it will be used for the remainder of the discussion. Although it was argued above that the dice coefficient gives a better indication of the actual amount of overlap between models, the absolute value of the similarity does not play any significant role for the cluster analysis which follows.

The joining or tree clustering method uses the dissimilarity between the various models, to analyse step-by-step how a family or relatedness tree of models could be constructed. Although there are a number of different clustering algorithms available, a simple, non-weighted clustering algorithm was used here whereby a (newly formed) cluster has the same weight as a single cluster in calculating the (next) new distance. Although this is not normal statistical practice, it shows the relative distances between references disciplines without being biased by the number of models in each reference discipline. For instance, when SR and OD and OB are joined, the resultant cluster should not receive a weight which is three times that of IN when it joins them later. Most statistical programs will output a joining tree (or dendrogram) with each join at a different horizontal (or vertical – depending on the orientation of the tree), but the tree has been re-drawn here (figure 1) to emphasize the family structure.

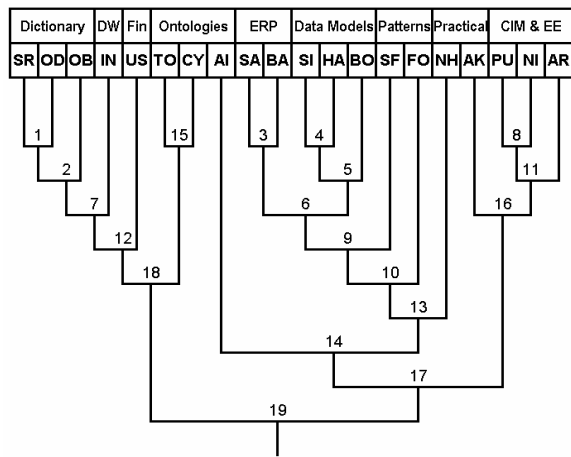


Figure 1: Similarity Dendrogram

This cluster analysis confirms, and is validated by, the similarity grouping between models from the same reference disciplines. The following is an attempt to verbalize the “cluster creation story” in a way which appears to make conceptual sense.

The first two steps combine the family models SR, OD and OB who owe their unusually close clustering by virtue of their design: they are all derived from the same base vocabulary list. Next is the closeness of the

two competing ERP systems: both are large, well-known and well-validated models, and it is not surprising to see them forming the first “real” model cluster. The next closest cluster (steps 4 and 5) combines the three data (library) models together: Silverston, Hay and BOMA. They follow a similar approach and cover the same domain, so it is not surprising, though a nice validation for the methodology, to see them cluster together so neatly. Another obvious cluster is formed by CIM/EE (Computer Integrated Manufacturing & Enterprise Engineering) reference discipline based Purdue, Nippon and ARRI.

A next step occurs where the ERP and data model clusters are joined: from a conceptual point of view; these can indeed be considered to be two disciplines that are fairly closely related. This ERP-Data Modelling cluster is subsequently joined by the two patterns-based models, which conceptually makes a lot of sense. The dictionary cluster expands by absorbing the data-warehousing Inmon model (which is indeed not much more than a vaguely structured set of terminology) as well as the USB financial model. (The Ottawa dictionary was very financially oriented.) As seen above, the TOVE and CYC models also form their ontology cluster, being very lexically oriented, which eventually joins the enlarged dictionary cluster. At this stage, any similarity between the clusters and remaining models has become fairly small.

Generally, the overall shape of a dendrogram can be described as either a black hole or a planetary system. In this particular case, the planetary system best describes the model similarity dendrogram, although the two clusters formed in stages 1 and 6 act as small black holes for a while, slowly but systematically absorbing their neighbours.

7.3 Three-dimensional surface plot of similarity metrics

In a further attempt to visualize the entire table of similarity metrics, Figure 2 shows a three-dimensional surface plot of the table containing the dice similarities, but with the models ordered in the same way as the hierarchical tree. The various peaks highlight clusters of similar models. The two guiding interpretation principles are: the higher the peak, the more similar the models are; and the broader or wider the peak, the more models share the similarity.

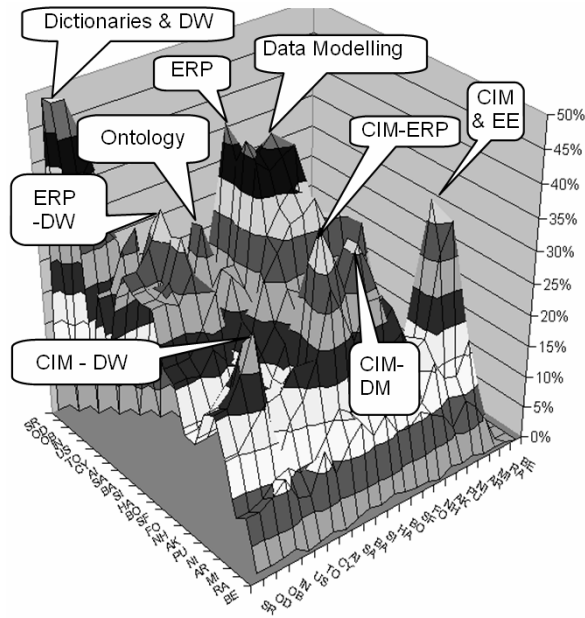


Figure 2: Visualization of Model Clusters

This grouping of models according to cluster closeness confirms the validity of the clustering according to reference discipline: most of the highest peaks are located towards the back, along the diagonal line of the plot. Note that the other, symmetrical back-half of the plot has been removed to reduce clutter. The back peaks representing the dictionaries (OD, OB, SR; top truncated at 50%), ontologies and CIM & EE stand out clearly, with the most prominent and largest “triple peak” mountain at the centre back representing ERP and Data Modelling.

However, this plot also highlights some peaks which are off the diagonal line and reveals other similarities between disciplines which were not evident in the clustering procedure (which gives an essentially one-dimension approach.) For example, although Inmon’s (data warehouse) model is mathematically closest to the Ottawa model (39-42%) and part of that peak, going along its ridge from left-back to centre-front the prominent sub-peaks with the ERP and CIM/EE models are encountered. Similarly, if we follow the ridge line of the CIM/EE models (right-back to centre-front), the almost equally tall sub-peaks representing their similarity with the DM (Data Modelling), ERP and DW (Inmon) models are very prominent.

8. Conclusion

An objective metric to measure the degree of semantic overlap between models was proposed. The fairly straightforward metric uses word labels

associated with each model element but attempts to compensate for synonymy and hypernymy using the WordNet linguistic resource. To aid the comparison of several models simultaneously, visualisation using dendrogram and 3D surface plots was explored.

A measure of validity of the metric was demonstrated by illustrating the approach on a database of fairly large real-world enterprise models. Although the analysis was restricted to entities in data models, in principle the metric can just as easily include other model elements (relationships and attributes). It was found that, despite the relative simplicity of the metric, in practice it performed fairly well for large models and the results clearly identified those models known to be closely related by practitioners correctly.

It is hoped that some or all of suggested metrics and analysis techniques (or refined versions thereof) will be found soon in CASE, enterprise architecture and business modelling tools soon to facilitate the comparison or integration of models.

Future research should further validate and refine the approaches presented here. In particular, the applicability to dynamic models should be investigated. An independent validation in a different domain area and comparison with expert opinion would be recommended. Another remaining challenge is to incorporate structural model information more fully into the similarity metrics.

9 References

- [1] Chen-Burger, Y.-H. *KBST: A Support Tool for Business Modelling in BSDM*. MSc. Thesis, Department of Artificial Intelligence, University of Edinburgh, 1994.
- [2] Chen-Burger, Y.-H.; Robertson, D. & Stader, J. “Formal Support for an Informal Business Modelling Method.” *Proceedings of the Tenth International Conference on Software Engineering and Knowledge Engineering (SEKE’98)*, 18-20 Jun 1998.
- [3] Noy, N.F. & Musen, M. A. “PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment.” *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, Austin, Texas, 2000, pp. 450-455.
- [4] Resnik, P. “Using Information Content To Evaluate Semantic Similarity In A Taxonomy.” *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, Montreal, Canada, 1995, pp.448-453.

- [5] Dastani, M. & Indurkha, B. "An Algebraic Approach to Similarity and Categorization." *Interdisciplinary Workshop On Similarity and Categorization*, Edinburgh, Scotland, 1997.
- [6] Lin, D. "An Information-Theoretic Definition Of Similarity." Proceedings of the *Fifteenth International Conference on Machine Learning*. San Francisco, California: Morgan Kaufmann, 1998, pp. 296-304.
- [7] Chen, Z. and Zhu, B. *Some Formal Analysis of the Rocchio's Similarity-based Relevance Feedback Algorithm*. Technical Report CS-00-22, Dept. of Computer Science, University of Texas-Pan American, Mar 2000.
- [8] Lee, L. *Similarity-Based Approaches to Natural Language Processing*. PhD thesis, Harvard University, 1997. (also Technical Report TR-11-97).
- [9] Chen, C. "Structuring and Visualising the WWW by Generalised Similarity Analysis." Proceedings of the *Eighth ACM Conference on Hypertext*, Apr 1997, pp. 177-186.
- [10] Aljlal, M. & Frieder, O. "Effective Arabic-English Cross-Language Information Retrieval via Machine-Readable Dictionaries and Machine Translation. *Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management*, Atlanta, Georgia, USA, November 5-10, 2001.
- [11] Van Belle, J.P. *A Framework for the Analysis and Evaluation of Enterprise Models*. Doctoral Thesis, University of Cape Town, 2003.
- [12] Honkela, T. *Self-organizing Maps in Natural Language Processing*. Doctoral Thesis, Helsinki University of Technology, 1998.
- [13] Scheer, A-W. *Business Process Engineering. Reference Models for Industrial Enterprises*. Berlin: Springer-Verlag, 1998 (2nd Ed.).
- [14] Perreault, Y. & Vlastic, T. *Implementing Baan IV*. Indianapolis, Indiana: Que, 1998.
- [15] Marshall, C. *Enterprise Modelling with UML. Designing Successful Software Through Business Analysis*. Reading, Massachusetts: Addison-Wesley, 2000.
- [16] Fowler, M. *Analysis Patterns*. Reading, Massachusetts: Addison-Wesley, 1997.
- [17] Hay, D.C. *Data Model Patterns*. London, UK: Dorset House, 1996.
- [18] Silverston, L.; Inmon W.H. & Graziano, K. *The Data Model Resource Book: A Library of Universal Data Models For All Enterprises*. New York: J. Wiley Computer Publishing, 2001.
- [19] Williams, T.J. (ed.). *A Reference Model For Computer Integrated Manufacturing (CIM). A Description from the Viewpoint of Industrial Automation*. CIM Reference Model Committee, The Instrument Society of America, Research Triangle Park, North Carolina, 1991 (2nd ed).
- [20] Uschold, M.; King, M.; Moralee, S. & Zorgios, Y. "The Enterprise Ontology." *The Knowledge Engineering Review*, Vol. 13 (1998).
- [21] Reyns, C.; Jorissen, A. and Vanneste, J. *Inleiding tot Accountancy*. UFSIA Universitaire Reeks Economie, Antwerp, Belgium, 1994.
- [22] Van Belle, J.P. *The USB Growth Model*. M.B.A. Thesis, University of Stellenbosch, 1988.
- [23] Miller, J. G. *Living Systems*. New York: McGraw-Hill, 1978.
- [24] Agirre, E.; Ansa, O.; Hovy, E. and D. Martinez. "Enriching Very Large Ontologies Using the WWW." Proceedings of the *First Workshop on Ontology Learning OL'2000*, Berlin, Germany, 25 Aug 2000.
- [25] Swartout, B.; Patil, R.; Knight, K. & Russ, T. "Toward Distributed Use of Large-Scale Ontologies." *Proceedings of the Tenth Workshop on Knowledge Acquisition for Knowledge-Based Systems*, Banff, Canada, Nov 1996.
- [26] Bisson, G.; Nedellec, C. & Canamero, D. "Designing Clustering Methods for Ontology Building - The Mo'K Workbench." Proceedings of the *First Workshop on Ontology Learning OL'2000*, Berlin, Germany, 25 Aug 2000.
- [27] Maedche, A. & Staab, S. *Comparing Ontologies – Similarity Measures and a Comparison Study*. Internal Report 408, Institute AIFB, Karlsruhe University, 2001.
- [28] Isaacs, J. D. & Aslam, J.A. *Investigating Measures for Pairwise Document Similarity*. PCS-TR99-357, 1999.
- [29] Ibrahimov, O.; Sethi, I. & Dimitrova, N. "Novel Similarity Based Clustering Algorithm for Grouping Broadcast News." *Proc. of SPIE Conf. 'Data Mining and Knowledge Discovery: Theory, Tools, and Technology IV'*, Vol. 4730, April 1-4, 2002, Orlando, Florida, pp. 394-304.
- [30] Duch, W. "Similarity Based Methods: A General Framework For Classification, Approximation And Association." *Control and Cybernetics* Vol. 29, No. 4 (2000).