Semantic Integration of Conceptual Models

Luís Costa[1(✉)], Cristóvão Sousa[1,2], and Carla Pereira[1,2]

[1] CIICESI-ESTG, Politécnico do Porto, Felgueiras, Portugal
{8050120,cds,csp}@estg.ipp.pt
[2] INESC TEC, Porto, Portugal

Abstract. In a collaborative conceptualisation process, the existence of several solutions for a given domain is a very common problem. Given this, specialists must reach a consensus on the concepts that will encompass the final solution. Therefore, this work aims to provide a tool for the integration of conceptual models in order to help specialists during the negotiation phase of developing the final shared model. This approach analyses the concepts of two models and shows the similar concepts to the specialists. The semantic similarity is obtained after three stages, namely: normalization, syntax analysis and semantic analysis. To evaluate the proposed approach, the values of precision and recall measures were calculated in two practical application scenarios. The obtained results proved to be better when compared to the existing tools when applied to semi-formal models (conceptual maps), and very close to the best tools focused on formal models (ontologies) integration.

Keywords: Semantic similarity · Conceptual modelling · Collaborative conceptualisation · Domain experts

1 Introduction

Nowadays, most of organisational activities are knowledge-intensive and carried out in a collaborative way. Providing knowledge-intensive support to intra and inter-organisational business processes, requires information management strategies based on domain experts' knowledge sharing practices. Those strategies typical include activities such as conceptualisation, representation, use and reuse of artefacts, able to handle the informational needs, related to the organisational activities. The design of such semantic artefacts still a challenge since they must be addressed in early stages of conceptualisation and involving domain experts. Within collaborative conceptualisation processes, the way conceptual modelling activities are performed and managed, has direct impact on the knowledge representation expressivity and consequently on the common understanding of the domain. Conceptual modelling emerges as a form of knowledge representation, since it establishes a network of concepts and conceptual relations for a given domain. In a process of collaborative conceptualisation involving several group of experts, more than one solution (conceptual model proposal) may emerge, and it is necessary to agree on which concepts and relations will be used in the shared final model [1]. Hereupon, the way integration of conceptual models is conducted, is critical.

Á. Rocha et al. (eds.), *Recent Advances in Information Systems and Technologies*,
Advances in Intelligent Systems and Computing 569, DOI 10.1007/978-3-319-56535-4_35

Although the literature is mature in terms of the integration (matching and merging) of formal knowledge representation models (e.g. ontologies), it is still incipient in what regards the integration of semi-formal models (e.g. concept maps). In this paper, we present an approach to support domain experts on reaching consensus around the result of the domain conceptualisation, providing the appropriate means to discover the best candidates (concepts and relations) to the shared model, based on an hybrid approach combining both syntactic and semantic measures, focused on how relations among concepts were defined.

2 Domain Knowledge Reuse in Collaborative Settings

A collaborative conceptualisation process (CCP) is the set of activities, involving a group of experts in the creation of conceptual representations, depicting a common view of a domain. Compared to an individual conceptualisation process, the CCP adds a set of social activities, such as conceptual negotiation and practical management activities [2]. Besides, a CPP encloses itself a collaborative learning process [3]. The typical result of a CPP is a semi-formal conceptual representation in the form of concept - relationship - concept (CRC), similar to a concept map [4]. Indeed, Maria et al. [5] and Basque and Lavoie [6] reinforce the importance of using concept maps in a collaborative environment for knowledge creation. The authors also conclude that there is evidence that collaborative concept maps, when compared with individually constructed conceptual maps, have a better quality of construction, benefiting creation, knowledge sharing and learning [6]. If, on one hand, CPP assures the definition of a reliable and useful information model that might be at the basis of a knowledge management system. It should, on the other hand, ensure the reusability of the generated semantic artefacts. In collaborative environments, the degree of knowledge reusability depends on: (i) how well defined the knowledge structures are, regarding their basic constructs and representation format (or form), and; (ii) the extend to which the conceptual representations are agreed [4]. In both cases, the semantic integration phase of CPP play an important role. It contributes to the discovery of the conceptual representations that establish a shared view of a domain and, maintains the basic semi-formal structure of the different conceptual proposals.

3 Integrating Conceptual Models

Currently, model integration tasks are closely related to the area of ontology matching. Most of the available tools are designed to support the identification of alignments[1] between ontologies. The Ontology Alignment Evaluation Initiative[2] (OAEI) organises, annually, an event aiming at evaluate the alignment results of several tools, in the scope of ontology matching. According to the results of the 2015 edition, the AML, Mamba, LogMap-C, LogMap, XMAP, GMap, DKP-AOM and LogMapLite were the tools with the best performances [7]. An extended comparison of the tools used in OAEI, can be

[1] Alignment consists of a list of matches containing elements identified as similar.
[2] http://oaei.ontologymatching.org.

found in [8, 9]. In general terms, the tools for ontologies' integration operate using the semantic information that characterizes an ontology, that is, the classes, data properties, object properties, instances, unions, and disjunctions. In addition to these elements, they use inference mechanisms as a way to increase the accuracy of the analysis. However, existing inference mechanisms are optimized for ontological models, being therefore almost exclusive to formal ontologies' integration tools. Thus, it seems obvious the infeasibility to apply these tools in the scope of semi-formal conceptual models. Considering the matching of lightweight ontologies (also called conceptual ontologies or semi-formal ontologies), the S-Match tool presents three different forms of correspondence analysis: (a) basic semantic matching; (b) minimal semantic matching and; (c) structure preserving semantic matching (SPSM) [10].

The algorithm for basic semantic matching determines the degree of similarity, considering the terms that name the concepts, the position of the concept in the tree (converted from the lightweight ontology) and the relations between the concepts in the tree. Minimal semantic matching results in a smaller list of matches when compared to basic semantic matching, which facilitates domain experts' interpretation of the results. SPSM is a variant of the basic semantic matching algorithm, but with the difference of preserving the structure of the concepts under analysis. This variant focus on a concept-based analysis and it only considers elements at the same structural (e.g., generic concepts with generic concepts and specific concepts with specific concepts) [10]. Moreover, it does not consider associative relations that might exist on the lightweight ontology, reducing the ontology to a simple tree. Together with S-Match, CmapTools[3] is the tool more closely related to the approach discussed in this paper, from the artefact point of view. CmapTools allows a comparison between models and suggests similar concepts. This is accomplished by a syntactic analysis of the terms that name the concepts, and by the analysis of the possible synonyms of a concept using WordNet [11].

4 Expert Centric Approach to Conceptual Models Integration

To assist specialists in the integration of conceptual models, this paper presents an interactive and iterative approach for calculating similarity between the elements of a conceptual model. It is interactive because it allows the involvement of the specialists [9] and iterative considering the progressive and incremental nature of the integration process, allowing the users to monitor the actions carried out throughout the integration process. The approach comprises three phases: normalization, syntactic analysis and semantic analysis (Fig. 1). The first phase consists of a models preparation stage for the aftermost syntactic and semantic phases. Briefly, it comprises computational linguistic analysis applied to the terms that name concepts and relations together with a categorization of relations based on an existing ontology of relations. In the second phase, syntactic measures are used to calculate the similarity between concepts, considering only the lexicon. In the last phase, semantic measures are applied to allow a more comprehensive analysis of similarity, considering also the conceptual structure of the

[3] http://cmap.ihmc.us/cmaptools.

models, the positioning of the concepts in a taxonomy, and the information extracted from an existing *corpus*[4] associated to a concept or a model. The process ends when the domain experts agree upon the resulting model.

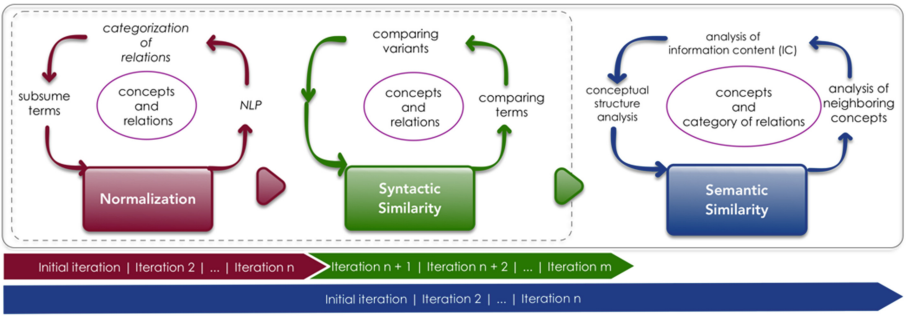

Fig. 1. Conceptual view of the process for the calculation of semantic similarity.

4.1 Phase 1: Normalization

The transformation of terms that define concepts and relations is performed, regardless of context, using natural language processing (NLP) mechanisms. By applying NLP, it is intended to eliminate linguistic variations of the concepts that can influence the results. In this case, we use stemming[5] algorithms and a list of stop words[6].

The categorization of the relations within a model is performed using the Conceptual Relations Reference Model (CRRM) ontology proposed by Sousa [4]. This categorization activity will allow a standardization of the interpretation of conceptual structures composing the conceptual models [12], and thus, enable a semantic analysis of the conceptual structure (in phase 3), beyond just the analysis of pairs of concepts.

Regarding the taxonomic relations, that is, is-a relations or generic-specific relations, the concepts will be subsumed to its generic concept, as a way to extend the analysis of similarity beyond the information of the concept itself, but also to the information of its generic [13]. In practical terms, and for calculations purposes, the specific concept is represented by its generic, however, all the information that characterizes the child concept (specific concept) is not discarded.

4.2 Phase 2: Syntactic Analysis

The syntactic analysis focus in the names and variants of the concepts of each model, to discover whether two concepts are close (identical) or distant syntactically.

[4] *Corpus* is a large and structured set of texts, used for statistical analysis, checking occurrences and validation of linguistic rules in a domain.
[5] Stemming: Process to reduce terms to their base language version [21].
[6] Stop Words: List of words of no relevance and that will not be considered [21].

The similarity value between two concepts, in this second phase, is obtained through the application of the following syntactic measures: Levenshtein, Jaro or MongeElkan (available in SimPack[7] and DKPro[8] APIs). Syntactically, the more common characters two concepts have, the more similar they will be.

In each iteration of the similarity process, the selection of the measure to be used is decision of the expert. Additionally, experts can define a minimum value to be considered as a valid match. This sensitivity parameter allows to exclude results with a degree of similarity below the certain value, filtering the list of final matches by eliminating matches with the lowest similarity value.

The simplicity of syntactic measures allows users to rapidly obtain a first list of matches with the duplicated concepts or concepts with a high probability of being similar. The major limitation of this type of measures consists both in the attribution of erroneous correspondences to the homograph terms (written in the same way, but with different meanings), and in the inability to detect correspondences between concepts written differently, but with the same meaning. For this reason, the approach proposed in this work also includes in its phase 3 semantic measures to overcome the syntactic limitations.

4.3 Phase 3: Semantic Analysis

In the third and last phase, semantic mechanisms are introduced to detect new matches and overcome the limitations of the second phase. With the semantic mechanisms, we intend to include the conceptual structures that composes a conceptual model.

The semantic similarity measures used in this third phase are grouped according to the characteristics of the conceptual models under analysis (Table 1). The quality of the similarity results is directly linked to the information that can be obtained from the conceptual models beyond the terms naming the concepts and relations.

For the structural analysis, based on the types of relations, the modified Dice measure will be used to support conceptual relations, henceforth called Dice - CRRM [4]. The resulting degree of similarity (Formula 1) depends on the number of common relations and the degree of relations of the concepts involved (number of relations to and from the concept).

$$sim(c1, c2) = \frac{2 \times nN\left(G_{c1}, H_{c2}\right)}{deg\left(G_{c1}\right) + deg\left(H_{c2}\right)} \tag{1}$$

Where $nN\left(G_{c1}, H_{c2}\right)$ is the number of relations between the concepts *c1* and *c2*, with the same category, and the degrees $deg\left(G_{c1}\right)$ and $deg\left(H_{c2}\right)$ correspond to the number of relations that link the concepts *c1* and *c2*, respectively. The application of the Dice-CRRM measure is exemplified in Fig. 2.

In the example above, there are two common relations between a and a' (*R1* and *R3*). The degree (*deg*) of existing relationships in each concept (a and a') is 3 (both

[7] https://files.ifi.uzh.ch/ddis/oldweb/ddis/research/simpack/index.html.
[8] https://dkpro.github.io.

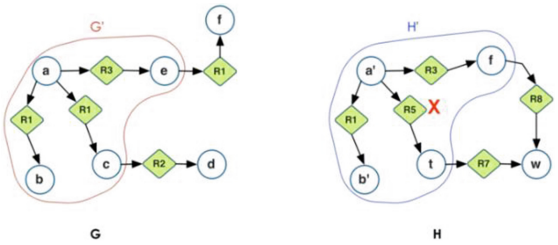

Fig. 2. Example of the application of the Dice-CRRM measure [4].

Table 1. Summary of the types of analyses and measurements used to calculate the similarity in the third phase.

Type of analysis	Information used	Description	Measure
Structural	Types of relations	Degree of similarity obtained according to the relation category assigned during the first phase	Dice-CRRM [4]
	Taxonomy	Similarity calculated from the distance between two concepts in the taxonomy	Wu e Palmer [14]
Information Content (IC)	Corpus	Similarity obtained by the computation of word co-occurrence in the *corpus* of the concepts under analysis	Cosine [15]

concepts have 3 directly linked relationships). Applying Formula (1), the similarity value between concept a and a' is:

$$sim(a, a') = \frac{2 \times 2}{3 + 3} \cong 0,67$$

If we are dealing with a taxonomy[9], the similarity value between pairs of concepts is directly related to their positioning within the taxonomy. The Wu and Palmer [14] measure, defined in Formula (2), considers the taxonomic distances between the concepts under analysis and the least common subsumer (LCS), and the distance between LCS to the root of the taxonomy.

[9] The taxonomy used, in each analysis, is constructed from the linking links (concepts integrated by the specialists during the iterations of the integration process of conceptual models) between the source model and the target model.

$$sim_{W\&P}(c_1, c_2) = \frac{2H}{N_1 + N_2 + 2H} \quad (2)$$

Where N_1 and N_2 are the is-a relationship numbers from the concepts *c1* and *c2*, respectively, up to LCS, and H is the number of is-a relations from LCS to the root of the taxonomy.

In the example presented in Fig. 3, it is intended to calculate the similarity between the concepts *A2* and *B3*, considering that the concepts *A1* and *B1* were "integrated" by the expert in an earlier iteration of the integration process. Thus, LCS (A2, B3) = (A1/B1) translates into the following similarity value:

$$sim_{W\&P}(A_2, B_3) = \frac{2}{1 + 1 + 2} = 0,5$$

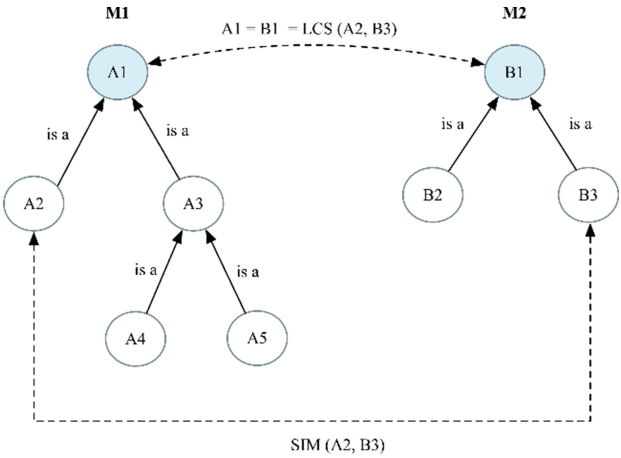

Fig. 3. Example of application of the taxonomy measure Wu and Palmer.

In the IC analysis, the Cosine measure [15] is used to determine the information set shared between two concepts considering the existence of a *corpus* created when the initialization of the conceptualization project [1]. The *corpus* of a concept might include the concept definition, related documents and variants (other terms to designate the concept). The more common elements (e.g. co-occurrence of words in the *corpus, variants similarity, etc.*) exist, the greater the similarity value.

5 Results and Conclusions

The approach proposed in this work was evaluated based on the quality of the results obtained in two scenarios. The first scenario consists of the use of the dataset of the conference domain, made available by OAEI 2015. In the second scenario, two conceptual models

were used from a collaborative conceptualisation process carried out in the scope of the Forsys[10] project.

The quality of the results were evaluated through the *precision* and *recall* measures [16]. These measures, widely used in ontology matching, allow to calculate, based on the reference results, the number of correct (True Positives - TP), incorrect (False Positive - FP) and not retrieved (False Negatives - FN) matches.

The *precision* measurement evaluates the ratio between the correct matches (TP) and total matches (TP + FP). Provides an indication of how many matches marked by the tools are indeed relevant.

The *recall* measure evaluates the ratio between the correct matches (TP) and total expected matches (TP + FN). This measure indicates how many relevant matches were marked in the alignments.

5.1 Scenario 1

For this scenario, the dataset composed of ontologies describing the domain of the conference organization, together with the respective results of the tools that participated in this initiative was be used [7]. Since the approach proposed in this work considers semi-formal models (concept maps), it was necessary to convert the ontologies, present in the dataset, into concept maps. In this transformation, a set of information that can be obtained from an ontology and its conceptual model was established. Considering the literature [17–20], an owl parser was implemented, using the similarities found between the elements of the ontologies and the elements of the concept maps.

Table 2 depicts the precision and recall values obtained by our approach (SimSemantica) in comparison to the reference alignments in the M1 modality (only contains classes).

Table 2. Results obtained in scenario 1.

Algo.	Prec.	Rec.
Reference	1,00	1,00
SimSemantica	**0,80**	**0,56**
AML	0,83	0,70
CroMatcher	0,72	0,51
DKP-AOM	0,84	0,59
GMap	0,76	0,71
JarvisOM	0,88	0,44
Lily	0,59	0,63
LogMap	0,83	0,54
LogMapC	0,84	0,52
LogMapLite	0,84	0,54
Mamba	0,84	0,66
RSDLWB	0,88	0,53

[10] http://www.cost.eu/COST_Actions/fps/FP0804.

Algo.	Prec.	Rec.
ServOMBI	0,56	0,44
XMAP	0,86	0,63
S-Match	0,39	0,30

5.2 Scenario 2

In this second scenario, two models were used from a collaborative conceptualisation process carried out in the scope of FORSYS project [4]. Additionally, the shared models and the integration decisions of the domains experts, gathered during the process, were also provided. The result of the integration is here considered as a reference, that is, the final solution expected by the integration tools. From this reference result, the precision and recall values are calculated, as described in the previous scenario. The alignment obtained by the proposed approach (**SimSemantica**) is compared to the result of the CmapTools (tool that is also directed to conceptual models). In this scenario, an automatic analysis was performed, without intervention of the specialists, and an iterative analysis, with the involvement of the specialists (Table 3, test 1 and test 2, respectively).

Table 3. Results obtained in scenario 2.

Algo.	Test 1		Test 2	
	Prec.	Rec.	Prec.	Rec.
Reference	1,00	1,00	n/a	n/a
SimSemantica	**0,88**	**0,88**	**0,80**	**1,00**
CmapTools	0,43	0,38	n/a	n/a

In comparative terms, the results obtained by SimSemantica clearly outperform the values of the CmapTools. In addition, in test 2, it is possible to check the usefulness of the interactive and iterative components, obtaining all expected matches (*recall* 100%), unlike CmapTools, which does not even consider the involvement of specialists.

6 Conclusions

The conceptual integration approach discussed in this paper, revealed to be highly flexible and, proved its usefulness considering the interesting results obtained, both in a scenario of integration of formal models (ontologies) and in a scenario of integration of semi-formal models (concept maps). Comparing the S-Match tool (also aimed at semi-formal models), the SimSemantica approach exibit better results. Regarding AML tool (the best for formal models), only a few tenths ahead of the "SimSemantica" approach discussed here. This is a clear indicator of the added value that this approach offers, both in the analysis of formal models and in semi-formal models. In the second scenario, the quality of the results achieved is even more evident, with 88% accuracy and recall, compared to the 43% accuracy and 38% recall obtained by the CmapTools.

In future works it is intended to use external resources in the analysis of similarity and to include mechanisms to guarantee the integrity of the relations of the merged

concepts. According to the results obtained, it is planned to include this approach and its services within a broader collaborative conceptual modelling environment.

Acknowledgments. This work is financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project «POCI-01-0145-FEDER-006961», and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013.

References

1. Pereira, C., Sousa, C., Lucas Soares, A.: A socio-semantic approach to collaborative domain conceptualization. In: Meersman, R., Herrero, P., Dillon, T. (eds.) OTM 2009. LNCS, vol. 5872, pp. 524–533. Springer, Heidelberg (2009). doi:10.1007/978-3-642-05290-3_66
2. Sousa, C.: Discussing and Collaborating Through Concepts: the Concept me Approach (2007)
3. Novak, J.D., Cañas, A.J.: The Theory Underlying Concept Maps and How to Construct and Use Them (2008)
4. Sousa, C.: Collaborative knowledge representation processes and techniques to support domain experts in conceptual modeling. Universidade do Porto (2015)
5. Maria, T., Dimitris, P., Garifallos, F., Athanasios, G., Roumeliotis, M.: Collaboration learning as a tool supporting value co-creation. Evaluating students learning through concept maps. Procedia Soc. Behav. Sci. **182**, 375–380 (2015)
6. Basque, J., Lavoie, M.-C.: Collaborative concept mapping in education: major research trends. In: Proceedings of the 2nd Second International Conference on Concept Mapping (2006)
7. Cheatham, M., et al.: Results of the Ontology Alignment Evaluation Initiative 2015. Ontol. Alignment Eval. Initiat., pp. 61–100, March 2016
8. Nentwig, M., Hartung, M., Ngonga, A., Rahm, E.: A survey of current link discovery frameworks. Semant. Web – Interoperability, Usability, Appl. J., 1–17 (2015)
9. Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges. IEEE Trans. Knowl. Data Eng. **25**(1), 158–176 (2013). Inst. Electr. Electron. Eng.
10. Giunchiglia, F., Autayeu, A., Pane, J.: S-Match: An open source framework for matching lightweight ontologies. Semant. Web **3**(3), 307–317 (2012)
11. Cañas, A.J., et al.: CmapTools: a knowledge modeling and sharing environment. In: Concept Maps: Theory, Methodology, Technology, Proceedings of the First International Conference on Concept Mapping, vol. 1(1984), pp. 125–135 (2004)
12. Sousa, C.D., Soares, A.L., Pereira, C.S.: Collaborative conceptualisation processes in the development of lightweight ontologies. VINE J. Inf. Knowl. Manag. Syst. **46**(2), 175–193 (2016)
13. Hussain, M.M.: A study of different ontology matching system. Int. J. Comput. Appl. **37**(12), 10–16 (2012)
14. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, ACL 1994, pp. 133–138 (1994)
15. Ganesan, P., Garcia-Molina, H., Widom, J.: Exploiting hierarchical domain structure to compute similarity. ACM Trans. Inf. Syst. **21**(1), 64–93 (2003)
16. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg (2007)
17. Graudina, V., Grundspenkis, J., Milasevica, S.: Ontology merging in the context of concept maps. Appl. Comput. Syst. **13**(1), 29–36 (2012)

18. Graudina, V., Grundspenkis, J.: Concept map generation from OWL ontologies. In: Proceedings of the Third International Conference on Concept Mapping, Tallinn, Estonia, Helsinki, Finland, pp. 263–270 (2008)
19. Graudina, V.: Owl Ontology Transformation Into Concept Map. Comput. Sci. **34**, 80–92 (2008)
20. Dean, M.: Annotation classes: a structuring mechanism for owl ontologies. CEUR Workshop Proc., vol. 496, pp. 1–6 (2009)
21. Sousa, C., Pereira, C., Soares, A.: Collaborative elicitation of conceptual representations: a corpus-based approach. In: Rocha, Á., Correia, A., Wilson, T., Stroetmann, K. (eds.) Advances in Information Systems and Technologies. AISC, vol. 206, pp. 111–124. Springer, Heidelberg (2013)