SEMCON: Semantic and Contextual Objective Metric

Conference Paper · February 2015

DOI: 10.13140/RG.2.1.5162.6728

3 authors:

Zenun Kastrati
Norwegian University of Science and Technology
19 PUBLICATIONS **23** CITATIONS
SEE PROFILE

Ali Shariq Imran
Norwegian University of Science and Technology
47 PUBLICATIONS **103** CITATIONS
SEE PROFILE

Sule Yildirim
Norwegian University of Science and Technolo…
94 PUBLICATIONS **156** CITATIONS
SEE PROFILE

SEMCON: Semantic and Contextual Objective Metric

Zenun Kastrati, Ali Shariq Imran, and Sule Yildirim Yayilgan
Gjøvik University College, Norway
Faculty of Computer Science and Media Technology
Email: {zenun.kastrati, ali.imran, sule.yayilgan}@hig.no

***Abstract*—This paper proposes a new objective metric called the SEMCON to enrich existing concepts in domain ontologies for describing and organizing multimedia documents. The SEMCON model exploits the document contextually and semantically. The preprocessing module collects a document and partitions that into several passages. Then a morpho-syntatic analysis is performed on the partitioned passages and a list of nouns as part-of-speech (POS) is extracted. An observation matrix based on statistical features is then computed followed by computing the contextual score. The semantics is then incorporated by computing a semantic similarity score between two terms - term (*noun*) that is extracted from a document and term that already exists in the ontology as a concept. Eventually, an overall objective score is computed by adding contextual score with semantic score. Subjective experiments are conducted to evaluate the performance of the SEMCON model. The model is compared with state-of-the-art *tf*idf* and χ^2 (Chi square) using F1 measure. The experimental results show that SEMCON achieved an improved accuracy of 10.64 % over the *tf*idf* and 13.04 % over the χ^2.**

I. Introduction

Domain ontologies are a good starting point to model in a principled way the basic vocabulary - concepts of a given domain. However, in order for an ontology to be actually usable in real applications, it is necessary to enrich concepts in ontology with available lexical resources of this particular domain. Concepts enrichment means adding new concepts without dealing with their ontological relations and types. Moreover, the ontology structure will remain the same but its concepts will be enriched with their synonyms and homonyms.

Recently, the population of the ontology with lexical data known as onto-terminology [1] has been the subject of research. In this regard, researchers in [2] proposed a new approach named *Synopsis* to automatically building a lexicon for each specific term called criterion. The authors used the assumption that terms appearing closer to a given criterion are more correlated to this criterion. The correlation is simply computed by only counting the number of grammatical terms between a given term and the user criterion. An adaptation of this approach is presented by researchers in [3]. They used the same methodology to build automatically the lexicon of an ontology concept in contrast to building a lexicon for a term. In order to do this, they built an information retrieval system called *CoLexIR* which automatically identifies all parts of a document that are related to a given concept. The issue of enriching the ontology concepts is also treated in [4] where researchers proposed a new methodology to enrich the upper-level ontology SUMO (Suggested Upper Merged Ontology) with the lexical data from the WordNet lexical database.

These approaches, using the the co-occurrence of terms, take into account only the contextual aspects of the domain in their learning process and do not consider the semantics. Therefore, this paper proposes a new approach namely SEMCON, which combines the contextual information and semantic information in the learning process of enriching the ontology concepts. Furthermore, in addition to frequency of occurrences of common noun terms, new statistical features such as term's font size and term's font type are introduced in this paper to build the observation matrix.

The rest of the paper is organized as follows. Section II describes the proposed SEMCON model in detail. In section III, we describe the subjective and objective experiment and we compare the subjective results with the results obtained by SEMCON model. Lastly, section IV concludes the paper.

II. SEMCON

This section describes the proposed SEMCON model to enrich concepts *c* of a domain ontology with new terms *t*. The model, illustrated in Figure 1, consists of 4 modules which are explained in the following subsections.

A. Preprocessing

This module first collects a document and partitions that into subsets of text known as passages. Each passage is treated as independent document in this paper.

Then a morpho-syntatic analysis is performed on the partitioned passages and the potential terms obtained can either be a noun, verb, adverb or adjective. These are different parts-of-speech (POS) of a language. It is a well established fact that nouns represent the most meaningful terms in a document [6], thus the focus of this paper is on extracting only common noun terms t_n for further consideration.

B. Observation Matrix

The second module of SEMCON deals with the calculation of the observation matrix. The observation matrix is formed using the frequency of occurrences of each term t_n, their font type (*bold*, *underline*, *italic*), and their font size (*title*, *level* 1, *level* 2) as given in equation 1. Using of font type and font size of a term is inspired from the representation of tags in

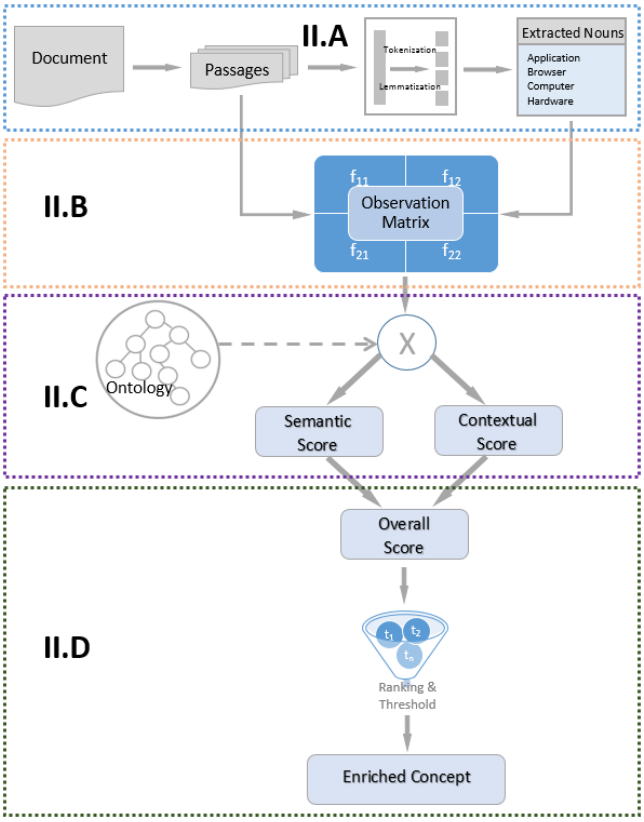

Fig. 1. Block diagram of SEMCON model.

the tag cloud. The font size and position of terms are found to be amongst the very important factors in the information finding process [5]. For instance, the bigger the font size is, the more important a term is in the given context.

$$O_{i,j} = \sum_{i \in t_n} \sum_{j \in p} (\alpha * Freq_{i,j} + \beta * Type_{i,j} + \gamma * Size_{i,j}) \quad (1)$$

where, t_n and p denotes the set of terms and set of passages respectively. α, β, γ are some constants set as 1 in our case. $Freq_{i,j}$ denotes the frequency of occurrences of term t_{ni} in passage p_j, $Type_{i,j}$ denotes term's font type t_{ni} in passage p_j, and $Size_{i,j}$ denotes term's font size t_{ni} in passage p_j.

We assumed that terms occurring in bold have more influence/effect on the readers than underline and than italic. According to this assumption, we computed the font type of a term t_n as given in equation 2.

$$Type(t_n) = 0.75 * B + 0.5 * U + 0.25 * I \quad (2)$$

Font size of a term t_n is calculated using equation 3.

$$Size(t_n) = 1.0 * T + 0.75 * L_1 + 0.50 * L_2 + 0.25 * L_3 \quad (3)$$

where T indicates title font size, L_1 indicates level 1 font size, L_2 indicates level 2 font size, L_3 indicates level 3 font size, B indicates the bold font type, U indicates underlined font type, and I indicates the italic font type.

The computation of each term's font size in the observation matrix is performed using the font sizes from a master slide in PowerPoint presentations where the level 1 font size is set to 28 pt, level 2 is set to 24 pt and level 3 is set to 20 pt.

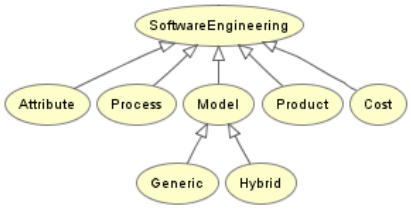

Fig. 2. Software engineering lightweight ontology

C. Contextual and Semantic Similarity

The observation matrix is used as input to compute the contextual and semantic similarity between two terms in order to match a term extracted from a passage with a concept in the ontology.

Term to term contextual distance, given in equation 4, is computed using the cosine measure in respect of passages.

$$S_{con}(t_{n1}, t_{n2}) = \frac{t_{n1} \cdot t_{n2}}{\| t_{n1} \| \| t_{n2} \|} \quad (4)$$

A term square matrix is used to store S_{con} values among all extracted terms t_n. This matrix will later be used in computing an overall correlation between a term extracted from a document and a concept in the ontology.

Further, we extract and use a subset of the terms t_n in order to extend the concept list of ontology. There maybe single label concepts in an ontology as well as compound label concepts. For single label concepts, we use only those terms from the term square matrix for which an exact term exists in the ontology. For example, for concept "Attribute" or "Generic" in the software engineering ontology shown in figure 2, there exists exactly a same term extracted in the term square matrix.

For compound label concepts, we use those terms from the term square matrix which are present as part of a concept in the ontology. For example, consider "SoftwareEngineering" as one of the compound label ontology concept, and the "Software" as one of the extracted terms from the document. Let "Program", "Design", "Development" be the highly correlated terms with the term "Software". In this case, the compound ontology concept "SoftwareEngineering" will be enriched with the correlation terms of the term "Software" namely with "Program", "Design", "Development".

The next step is the computation of the semantic similarity. The semantic similarity score, given in equation 5, is calculated using the Wu&Palmer algorithm [7] implemented in a freely available software package WordNet::Similarity [8].

$$S_{sem}(t_n, c) = \frac{2 * depth(lcs)}{depth(t_n) + depth(c)} \quad (5)$$

where t_n, indicates term extracted from document, c denotes term that already exists in ontology as a concept, *depth(lcs)*

indicates least common subsumer of term and concept label, *depth(t_n)* indicates the path's depth of term in WordNet::similarity and *depth(c)* indicates path's depth of concept label in WordNet::similarity.

D. Overall Score

The overall correlation of a term extracted from a document and a concept in the ontology is computed using the contextual and semantic score and it is given in equation 6.

$$S_{overall}(t_n, c) = w * S_{con}(t_n, c) + (1 - w) * S_{sem}(t_n, c) \quad (6)$$

where w is a parameter with value set as 0.5 in our case.

Finally, in order to obtain the terms which are more closely related to the ontology concepts, a rank cut-off method is applied to the terms t_n using a specified threshold. Terms which are above the threshold are considered to be the relevant terms for enriching the ontology concepts.

III. Experimental Procedures

To evaluate the performance of SEMCON, we have used PowerPoint presentations dataset from 5 different domains: Computer, C++ Programming, Database, Internet and Software Engineering. We were restricted to a maximum of 5 presentations with a limited number of slides (39 slides), due to subjective nature of the experiment.

The paper uses two approaches to evaluate the performance of the SEMCON. The first one is subjective evaluation and the second one is the objective evaluation. The results from software engineering domain are presented in this paper.

A. Subjective evaluation

To compare term to concept correlation obtained from the SEMCON, an online survey based experiment is conducted. The subjective survey was carried out by publishing online a questionnaire to 10 subjects. The subjects were all computer science PhD students and postdocs at the Gjøvik University College. They were asked to select 5 closely related terms from a list of terms for each concept, for 5 different domains, starting from the most relevant term as their first choice, the second relevant term as the second choice and so on.

From the subjective survey, a single score, for each selected term, is calculated using the Borda count method. Borda count, given in equation 7, is an election method used to determine a winner from a voting where voters rank the candidates in order of preference [9].

$$BordaCount(t_n) = \sum_{i=1}^{m} [(m + 1 - i) * freq_i(t_n)] \quad (7)$$

where *BordaCount(t_n)* of a given term t_n is calculated by a total sum of the weights of the frequencies *$freq_i(t_n)$*. *$freq_i(t_n)$* is the frequency of term t_n chosen at Position *i*, and *m* is the total number of possible positions, in our case *m* = 5.

The scores from the Borda count are then sorted to obtain the top 'n' terms, giving us the refined list of the highest scoring terms. For our experiment, we set n = 10, and this gives us the top 10 terms as shown in Table I. The term "Waterfall" has the highest Borda count value cause this term is selected by most of the subjects as the closest term for term "Generic".

TABLE I
BORDA COUNT OF SUBJECTS' RESPONSES FOR "GENERIC" CONCEPT.

Rank	Term	Borda Count
1	Waterfall	36
2	Model	16
3	Generic	10
4	Specification	10
5	System	10
6	Design	8
7	Transformation	7
8	Development	6
9	Phase	5
10	Formal	4

TABLE II
THE PERFORMANCE OF SEMCON

Concept	Precision (%)	Recall (%)	F1 (%)
Software	40.0	60.0	48.0
Cost	40.0	60.0	48.0
Product	40.0	60.0	48.0
Attribute	46.7	70.0	56.0
Process	60.0	90.0	72.0
Generic	60.0	90.0	72.0
Hybrid	60.0	90.0	72.0
Average	**49.5**	**74.3**	**59.4**

B. Objective evaluation

The second approach used to evaluate the performance of SEMCON is comparing the results obtained from the SEMCON with results obtained from the *tf*idf* and χ^2.

*tf*idf* is a mathematical algorithm which is used to find key vocabulary that best represents the texts by applying the term frequency and the inverted document frequency together [11].

The traditional *tf*idf* considers only the term to document relation and thus it is not appropriate for comparison as it is. Therefore, in order to take into account the term to term relation, cosine measure is used where the dot product between two vectors of *tf*idf* matrix reflects the extent to which two terms have a similar occurrence pattern in the vector space.

χ^2 is a statistical measurement which computes the degree of interdependency between any two terms [10]. The measurement is carried out by comparing the observation frequency with expected frequency.

We evaluated the performance of objective methods using the top terms scored by these methods. In order to do this, scores for the 10 top terms are taken as the ground truth, and they are compared with the top terms obtained by the objective scores. We used the top 15 terms as the refined terms list, and the effectiveness of objective metrics using the standard information retrieval measures are computed in order to compare with the subjective results. These measures are Precision, Recall and F1. Precision is the number of correctly retrieved terms, while recall is the number of retrieved terms. The F1 is considered as average of precision and recall.

Table II shows precision, recall and F1 results obtained from the SEMCON for software engineering concepts.

TABLE III
THE PERFORMANCE OF OBJECTIVE METHODS USING THE F1 MEASURE.

Concept	tf*idf (%)	χ^2 (%)	SEMCON (%)
Software	56.0	56.0	48.0
Cost	40.0	40.0	48.0
Product	64.0	56.0	48.0
Attribute	32.0	48.0	56.0
Process	72.0	48.0	72.0
Generic	48.0	64.0	72.0
Hybrid	64.0	56.0	72.0
Average	**53.7**	**52.6**	**59.4**

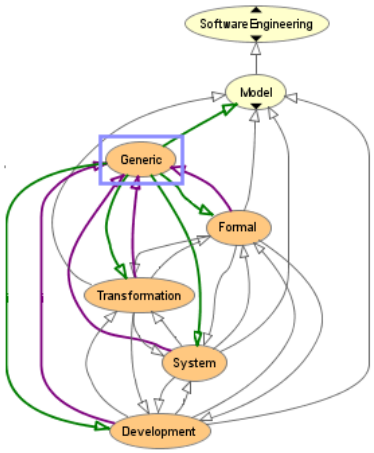

Fig. 3. "Generic" concept enriched with new terms

TABLE IV
F1 MEASURE FOR 5 DIFFERENT DOMAINS

Domain name	tf*idf (%)	χ^2 (%)	SEMCON (%)
Computer	44.44	40.89	49.78
C++_Programming	43.20	43.20	44.80
Database	40.00	32.10	41.00
Internet	49.14	41.14	45.71
SoftwareEngineering	53.71	52.57	59.43

The performance of SEMCON, in terms of F1 measure, is compared with the performance of *tf*idf* and χ^2. The comparison, depicted in Table III, shows that the SEMCON has achieved an improvement on finding the most related terms to enrich the concepts of an ontology, of 10.64 % over the *tf*idf* and 13.04 % over the χ^2. This improvement is achieved for all concepts excepts for "Software" and "Product". This may have happened due to the fact that the SEMCON, in contrast to *tf*idf* and χ^2, takes into consideration not only the frequency of occurrences of those terms in the corpus but also the semantics of those terms.

An example of a concept ontology enriched with new terms obtained by SEMCON is shown in Figure 3. The "Generic" concept of the software engineering ontology is enriched with new terms such as "System", "Development", "Formal" and "Transformation". These terms are amongst the top 10 terms selected also by subjects in the subjective experiment.

Finally, we evaluated the performance of the objective methods to a larger dataset comprised of lightweight ontologies from domains such as computer, database, internet and C++ programming (C++). The same experiment, as per Software Engineering ontology, was conducted. The obtained results in terms of F1 measure indicated in Table IV show that the SEMCON gives better results then both of the methods for all domains excepts for internet domain ontology. This may have happened due to the fact that subjects are making their selections based on descriptions provided under each concept on the questionnaire, when they were asked to select the 5 more closely related terms. Therefore, this causes the overall score to be mainly affected by the contextual score.

IV. CONCLUSION

This paper proposed a new objective metric namely SEMCON to enriching the domain ontology with new concepts by combining contextual as well as semantics of a term. The proposed method can be applied to any existing domain ontology for extending it with new concepts. The SEMCON takes into account the context of a term by first computing an observation matrix which exploits the statistical features. Currently three features - frequency of the occurrence of a term, term's font type and font size are used to compute observation matrix. These features can easily be extended based on the type of the document chosen. The future work may exploits further features for calculating observation matrix, and extracting candidate terms from multiple documents including word documents, audio and video files. We also plan to conduct further research to examine the contribution of the contextual and semantic scores in the overall score.

REFERENCES

[1] Roche, C., Calberg-Challot, M., Damas, L., and Rouard, P., "Ontoterminology - a new paradigm for terminology.", In Jan L. G. Dietz, editor, Proceedings of the International Conference on Knowledge Engineering and Ontology Development, Portugal, 2009.

[2] Duthil, B., Trousset, F., Roche, M., Dray, G., Plantié, M., Montmain, J., and Poncelet, P., "Towards an automatic characterization of criteria.", In: Proceedings of the 22^{nd} International Conference on Database and Expert Systems Applications, 2011.

[3] Ranwez, S., Duthil, B., Sy, M.F., Montmain, J., Augereau, P., and Ranwez, V., "How ontology based information retrieval systems may benefit from lexical text analysis.", In: New Trends of Research in Ontologies and Lexical Resources,(chapter 11), pp. 209-228, 2013.

[4] Niles, I., and Pease, A., "Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology", In Proceedings of the IEEE Conference on Information and Knowledge Engineering, 2003.

[5] Halvey, M.J., and Keane, M.T., "An assessment of tag presentation techniques", in Proceedings of the 16^{th} International Conference on World Wide Web, USA, pp. 13131314, ACM. 2007.

[6] Li, H., Tian, Y., Ye, B., and Cai, Q., "Comparison of current semantic similarity methods in wordnet", in Computer Application and System Modeling (ICCASM), vol. 4, pp. 4008-4011, 2010.

[7] Wu, Z., and Palmer, M., "Verb semantics and lexical selection", in Proceedings of the 32^{nd} Annual Meeting of the Associations for Computational Linguistics, pp. 133-138, 1994.

[8] Pedersen, T., Patwardhan, S., and Michelizzi, J., "Wordnet::Similarity - measuring the relatedness of concepts.", in Proceedings of 19^{th} National Conference on Artificial Intelligence, pp. 1024-1025, 2004

[9] Young, P., "Optimal voting rules.", The Journal of Economic Perspectives, vol. 9, pp. 51-64, 1995.

[10] Liu, J.N.K., He, Y-L., Lim, E.H.Y., Wang, X-Z., "A New Method for Knowledge and Information Management Domain Ontology Graph Model.", IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 43, 2013

[11] Sebastiani, F., "Machine learning in automated text categorization", ACM Computing Surveys (CSUR), vol. 34, no. 1, pp. 1 - 47, 2002.