Lucinéia Souza Maia – Universidade Federal de Ouro Preto, Brazil
Gercina Ângela de Lima – Universidade Federal de Minas Gerais, Brazil

A System for Specifying Semantic Relations for Knowledge Representation

Abstract:
Semantic relations are fundamental for understanding the nature of the connection between two concepts in a domain. This paper presents a model for extracting semantic relations for the representation of knowledge from academic documents in the context of the Portuguese language. A Web information system called Semantizar was developed to support the extraction of semantic relations from classificatory structures that represent specific academic documents. To evaluate the qualitative performance of Semantizar, a case study was carried out, which pointed to important contributions to research about semantic relations extraction. According to outcomes, when two concepts of a classificatory structure exist in a sentence, a semantic relationship between them can actually exist. Finally, it is concluded that this research is relevant because it brings important findings for the extraction of semantic relations for the knowledge representation of academic documents to be applied in the Brazilian scenario.

1.0 Introduction

Semantic relations are fundamental for understanding the nature of the connection between two concepts in a domain. According to Khoo and Na (2006) and Green, Bean e Myaeng (2011), concepts can be seen as blocks of knowledge, and relationships are links that connect and hold these blocks together within the structures of knowledge in people's minds.

Some factors can influence the specification of semantic relations, including language and culture. According to Khoo and Na (2006), it is difficult to analyze the meaning of concepts and their relations when they are taken apart from language because each language has its characteristics and is linked to the cultural factor. According to Khoo and Na (2006), it is difficult to analyze the meaning of concepts and their relations when they are taken apart from language because each language has its characteristics and is linked to the cultural factor. Therefore, in the literature review conducted by Maia (2018), was identified the lack of research concerning the theme in Brazil.

Subsequently, a Semantic Relations Extraction Model was formulated. This model's goal is to be a theoretical construction of a way of establishing relationships between concepts from academic document, in order to create semantic structures. The model was expanded to a computational prototype called Semantizar.

2.0 The Semantizar

The following procedures were observed in the development of Semantizar: (1) specification, (2) data modeling, (3) architectural design, and (4) prototype implementation in a Web system. Of these, the specification and the prototype implementation will be presented in this paper.

In the first stage of the development of Semantizar - specification -, was made the description of the algorithm for the Semantic Relations Extraction Model, which considers as inputs a classificatory structure and its respective academic document, from which the structure originated. The classificatory structure is broken down into concepts and the academic document is decomposed into sentences.

After this decomposition into concepts and phrases, the Semantizar scans all the phrases searching for pairs of concepts. In the first scan, the pair of concepts 1 and 2 is enhanced, then is checked their existence in each sentence until the last one, regardless of the discovery of a pair of concepts in a sentence. In this way, the search is done throughout the document. If the concept pair is found in a sentence, that sentence is highlighted, so that a manual check confirms or denies the existence of a semantic relationship between these concepts. Then, the Semantizar combines the concept 1 with all the other concepts of the classificatory structure, checking whether each combination exists in a sentence. Upon ending the combinations with concept 1, the Semantizar subsequently makes combinations with concept 2 and checks whether they exist in all sentences. In this way, the Semantizar makes combinations of pairs of concepts with all the concepts of the classificatory structure and checks the existence of each of these pairs in each sentence of the academic document, from the first to the last one, as illustrated in Figure 1.

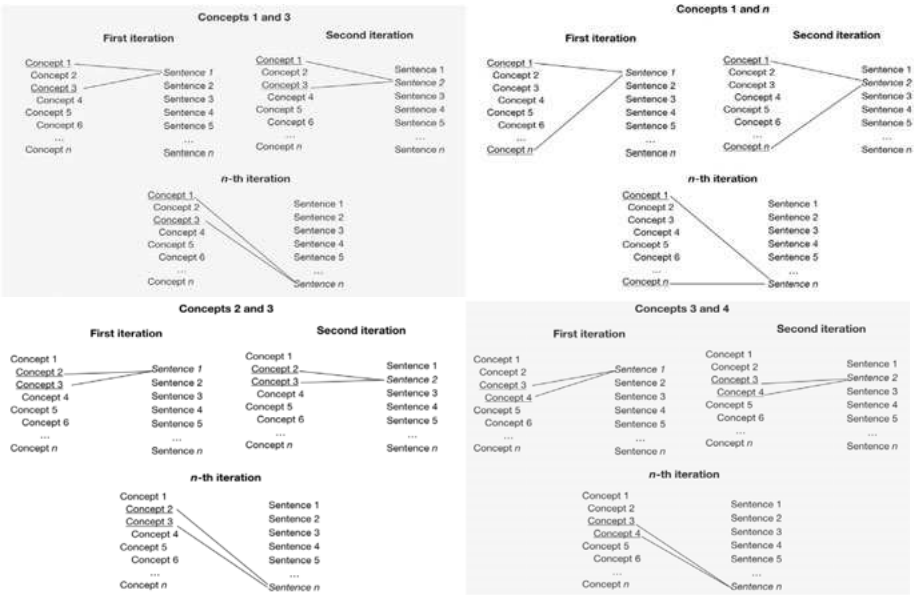

Figure 1: Search iterations of pairs of concepts in the sentences of the academic document.

The prototype implementation phase was divided into three activities: (1) data input, (2) reading and preparation and (3) extraction of semantic relations. The Figure 2 shows de initial interface of Semantizar prototype.

In the data input activity (1), the user informs the metadata of the academic document from which intends to extract semantic relations, and then sends the files of the publication and classificatory structure. The publication being a .pdf file and the classificatory structure being a plain text file with .txt extension.

Figure 2: The initial interface of Semantizar.

In the subsequent activity, reading and preparation (2), the Semantizar checks whether each term of the classificatory structure exists in the database. If the term does not exist, it is automatically registered by the system in way: The PHP programming language, chosen for the implementation of the model, allows text files to be converted into vectors. In this sense, the file that refers to the classificatory structure is automatically converted into a vector of terms, in which each line of the structure (which refers to a term) is transformed into a position of the vector. Therefore, the algorithm goes through each position of the vector, checking if the content of the position, which is the term of the classificatory structure, exists in the database; if it does not exist, the term is automatically registered by the system as a noun. This was established because, grammatically analyzing, the terms of the classificatory structure denote the nouns.

The second task of the reading and preparation activity is the preparation of the publication file (which is an academic document: a thesis or a dissertation) for the manipulation that will occur in the next activity of the prototype. Due to the programming language chosen for the implementation, it was necessary to convert the text in .pdf format to a temporary file in .txt format.

The activity of extracting semantic relations (3) is considered the most important, being the core of the Semantic Relations Extraction Model. It consists of two tasks: the first searches for pairs of terms from the classificatory structure in sentences of the publication to which the structure refers. The temporary file of the academic document created in the previous step is transformed into a string (variable that stores alphanumeric characters). Subsequently, this string is decomposed into a smaller string each time a period (.) is found in the file. In this way, each position of the vector is a sentence of the publication, separated by a period. Then, the sentence vector is scanned seeking terms of term vector in each sentence vector position. If a term is found in the sentence, the Semantizar goes through the other positions of the vector of terms checking if there is another term of the structure in the same sentence. In affirmative case, the sentence is taken to compose the interface created for the user to validate the semantic relationship.

The validation of the semantic relation is the second task of the relation extraction activity. If the user agrees that there is a semantic relation between the two concepts of

the classificatory structure found in a sentence in the publication by the Semantizar, (s)he is directed to semantic relation register interface.

In the register interface the user specifies the semantic relation according to judgment when analyzing the sentence. Besides that, the user determines the type of semantic relation, the inverse relation, if any, and points out the properties: symmetry and reflexivity. The transitivity property was not considered because it is understood that it applies to ternary relationships, which is not the case in this paper.

The types of semantic relations that appear in Semantizar were the result of a literature review by Maia, Lima and Maculan (2017), which elaborated a taxonomy with 63 types of semantic relations classified as hierarchical, equivalent and associative.

3.0 Case study

To evaluate the efficiency of Semantizar in the extraction of semantic relations, a case study was carried out. It was organized according to the Experimentation Process methodology proposed by Wohlin et al. (2014), which considers three stages: (1) definition of the case study, (2) planning and (3) operation.

In the first stage, definition of the case study (1), it was determined: (a) the object of the case study, which is the semantic relations; (b) the objective: to verify the efficiency of Semantizar; and (c) the context, which are theses and dissertations in the domain of Knowledge Organization and Representation.

After the case study definition phase, the second stage, planning (2), follows. According to Wohlin et al. (2014), this stage basically indicates "how" the case study will be conducted. Thus, in planning, it was established: (a) the sample and (b) the analysis of the data to be collected.

For the sample, was chosen the faceted structure of MHTX, by Lima (2004) and the thesis *Fatores Interferentes no Processo de Análise de Assunto: Estudo de Caso de Indexadores* (Interfering factors in the process of subject analysis: indexer case study), by Naves (2000). The MHTX is an in-context hypertextual navigation model to organize theses and dissertations, aiming to support the reading and retrieval of these documents in Digital Libraries of Theses and Dissertations. In the MHTX prototype three navigation tools was created: the expanded summary, the concept map and the faceted structure. In its implementation, the Naves's (2000) was used aforementioned thesis to instantiate the tools created. Among these tools, the faceted structure presents the characteristics of the desired sample for Semantizar.

As mentioned in the planning stage, in addition to defining the sample, the analysis of the data was determined. In this case, was decided to perform quantitative and qualitative analyses in order to identify: (I) the number of semantic relationships suggested by the application due to the amount of semantic relations that actually exist (this factor is important in order to evaluate whether the Semantizar has the potential to automatically extract semantic relations); (II) the concepts that are most likely to be semantically related (this parameter can point to the key concepts of the analyzed publication); (III) the characteristics of the semantic relations found; and (IV) the parallel between concept relationships in the original faceted structure and the resulting representation from Semantizar.

Following the case study process, the next step is the operation (3). This phase consists of the execution of the previously defined and planned case study (Wohlin et al.

2014). For this, three procedures were necessary. The first was preparation, which involved the clipping of the sample from the classificatory structure and the publication, selecting subjects from the Personality facet, as can be seen in Figure 3. In this case, the terms idea, thought and concept were broken down. Regarding the academic document, we decided to consider chapters 2, 3 and 4 of Naves's thesis (2000). This choice was due to the fact that these chapters constitute the conceptual definitions in the thesis in question. Figure 4 shows the thesis' summary, in which these chapters can be seen. We also decided to remove images from the chapters, since Semantizar cannot support image analysis. The second procedure, execution, comprised the processing of the sample snippets on Semantizar. Finally, the last procedure was the validation, in which a refinement and compilation of the data collected was made in order to avoid interferences in the data analysis and interpretation. Both in the validation of the data and in the results, the concepts were observed individually and with their pairs.

Personalidade [Entities]

- Autores
- Profissional da informação
  - Bibliotecário

(Pela natureza do seu trabalho)

  - Indexador

(Pela experiência)

    - Indexador experiente
    - Indexador pouco experiente
    - Indexador novato

(Pelo grau de conhecimento)

    - Especialização
    - Prática
- Conceito/Idéia/Pensamento
- Documento
- Texto

(Pela natureza do texto)

  - Narrativos
  - Informativo
  - Primário
  - Secundário
  - Hipertexto

(Pela estrutura)

  - Microestrutura
  - Macroestrutura
  - Superestrutura

Figure 3: Snippet of MHTX faceted structure

Figure 4: Fragment of the expanded summary of the Naves (2000)

4.0 Results

Figure 5 presents a conceptual map generated from the classificatory structure without Semantizar. In this Figure, it appears that the explicit semantic relations are due to the naming of the sub-facets used by Lima (2004) (see these sub-facets highlighted in Figure 3). Also, it is observed that there is a semantic relation between information professional (*Profissional da Informação*) and librarian (*Bibliotecário*) created from indentation that denotes a type of hierarchy. However, it was not possible to specify what the semantic relation really is.

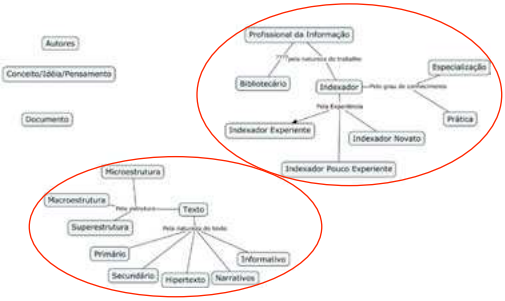

Figure 5: Clusters resulting from conceptual map of the MHTX faceted structure without Semantizar.

In the organization of the faceted structure in the concept map, the presence of two clusters is verified, as highlighted in Figure 5. The first group consists of concepts related to text (*texto*), and the other consists of concepts related to indexer (*indexador*). In clusters, objects belonging to a group are related to each other, however, they do not relate to concepts that are outside of their group. Therefore, in the conceptual map generated from the semantic relations found in Semantizar, there is a cohesion between all concepts in such way that these clusters are not possible to be obtained, that is, the concepts are all related to each other, as seen in Figure 6. With the use of Semantizar, it was possible to explicit 101 semantic relations.

5.0 Contributions

The relations between all concepts were possible with the support of Semantizar, which allowed the creation of a representation that covered every concepts of a semantically related classificatory structure. So, the user can view all possible relations between the concepts. Therefore, Semantizar achieved its goal of semantically enriching a classificatory structure.

The performance of the case study, within the scope presented, was possible due to the computational support of Semantizar. This task performed manually could demand time and effort on the part of the professional who executes it. In this sense, Semantizar facilitated the extraction and explanation of semantic relations, essential for the knowledge representation, semi-automating this task. Thus, the Semantizar contributed to the knowledge representation based on a classificatory structure, showing itself to be objective when detecting two concepts in a sentence from extensive text. The identification of concept pairs within sentences is one of the most laborious steps in the context that Semantizar was created to operate.

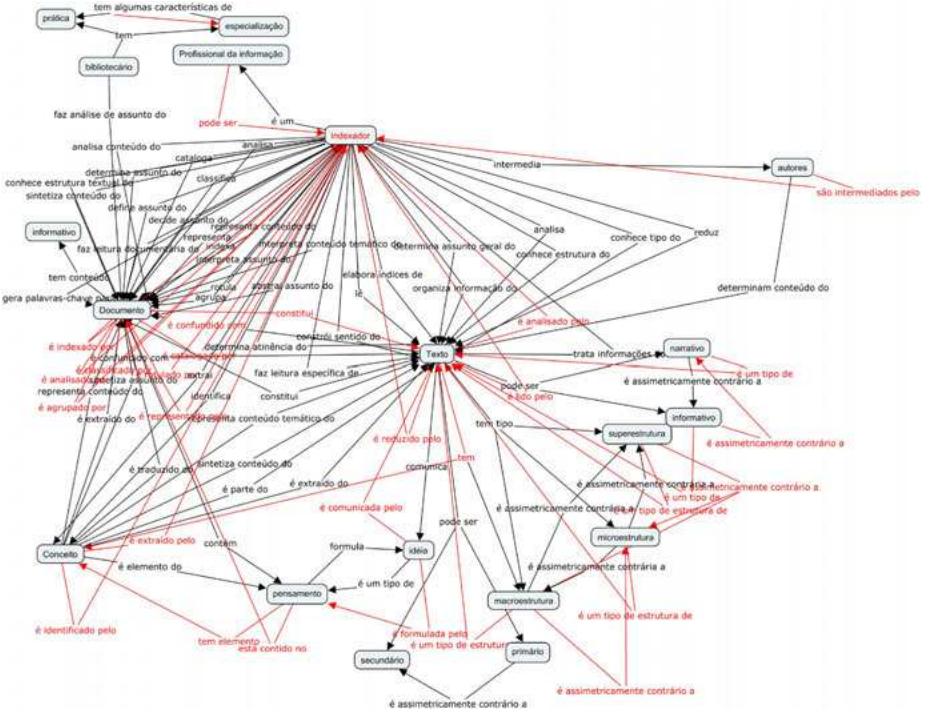

Figure 14: Conceptual map of the semantic relations established in the case study.

Therefore, through Semantizar, it was possible to establish 101 semantic relations, including inverse ones, in 53 different concept pairs, based on a classificatory structure containing 22 concepts. With that, it was possible to improve the semantics of the sample.

Another important contribution of Semantizar was that, in a way, it acted as a validating agent of the classificatory structure, as it indicated 199 signs of semantic relations in 131 sentences from the snippet of the academic document from which the classificatory structure originated. It is believed that this amount of evidence is an important indicator to suggest that the structure relevantly represents the academic document. Another important factor in this regard concerns the sentences in the academic document, as some of them encompassed more than a clue of semantic relation, which reinforced confidence in the classificatory structure.

In addition, it was possible, through Semantizar, to indicate the most important concept pairs for the academic document. During the analysis of the data, it was found the concept pairs that occurred most frequently were the most relevant in the represented domain. In the case of the sample, two pairs stood out: indexer and document (*indexador e documento*) and indexer and text (*indexador e texto*). Considering these pairs were extracted from the thesis entitled Interfering factors in the process of subject analysis: indexer case study, by Naves (2000), it can naturally be said, even without reading the entire content of the academic document, that such pairs are the most important in the

domain. Consequently, Semantizar also indicated the most important concepts for the academic document. In the same way as for the pairs of concepts, considering the concepts that occurred most frequently in the semantic relations, it can be said that the most important ones in Naves's thesis (2000) were indexer (*indexador*), text (*texto*) and document (*documento*), also taking into consideration that text and document were classified as almost synonyms, which also indicates coherence in determining this semantic relationship.

It was also found during the analysis that, many times, the inverse relations were not possible because the relations were indirect, the concepts existed: subject (*assunto*), content (*conteúdo*) and information (*informação*). Therefore, it was discovered that these concepts should compose the classificatory structure due to the fact that they repeatedly occurred and that they are representative for the domain. Similarly, it was noticed the concepts with the most evidences false in the context in which they were found were those that are routinely used in academic documents, such as concept (conceito), idea (ideia) and authors (autores). In this sense, a review of the classificatory structure to indicate whether they remain to represent the academic document was suggested. Thus, in pointing out these suggestions, it can be said that Semantizar can operate to refine the classificatory structure.

It was also noted that the extraction of concepts can be performed from lists of terms, as the inherent hierarchy of the classificatory structure was not decisive in Semantizar for the indication of the existence of a semantic relation.

Finally, the case study carried out made important contributions to research on the extraction of semantic relations. Those contributions are: 1) The presence of two concepts in a sentence is an indication of the existence of a semantic relation between these concepts; 2) A pair of concepts can have more than one semantic relation; 3) A pair of concepts can have the same semantic relation, even in different contexts; 4) The context, and the knowledge about it, is fundamental for determining the type and/or subtype of the semantic relation for the same pair of concepts with different relations; 5) The determination of semantic relations, in the way it was treated by Semantizar, depends on human interpretation; 6) The verbs are the main grammatical class for defining a semantic relation and; 7) Not all semantic relations can be explained.

References

Green, Rebecca, Carol A. Bean, and Sung Hyon Myaeng. 2011. *The Semantics of Relationships: An Interdisciplinary Perspective*. Dordrecht: Springer.

Khoo, Christopher S.G. and Jin-Cheon Na. 2006. "Semantic Relations in Information Science." *Annual Review of Information Science and Technology* 40, no. 1: 157-228.

Lima, Gercina Angela Borem de Oliveira. 2004. *Mapa Hipertextual (MHTX): Um Modelo para Organização Hipertextual de Documento*. Ph.D. dissertation. Belo Horizonte: Federal University of Minas Gerais.

Maia, Lucinéia Souza. 2018. *Extração e Explicitação de Relações Semânticas para a Representação do Conhecimento de Documentos Acadêmicos: Um Estudo de Caso a Partir de uma Estrutura Classificatória*. Ph.D. dissertation. Belo Horizonte: Federal University of Minas Gerais.

Maia, Lucinéia Souza, Lima, Gercina Ângela de, and Maculan, Benildes Coura Moreira dos Santos. 2017. "Taxonomia dos Tipos de Relações Semânticas para a Organização e Representação do Conhecimento: Uma Proposta a Partir da Literatura." *Tendências da Pesquisa Brasileira em Ciência da Informação* 10, no. 2.

https://revistas.ancib.org/index.php/tpbci/article/view/419/418

Naves, Madalena Martins Lopes. 2000. *Fatores Interferentes no Processo de Análise de Assunto: Estudo de Caso de Indexadores*. Ph.D. dissertation. Belo Horizonte: Federal University of Minas Gerais.

Wohlin, Claes, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. 2014. *Experimentation in Software Engineering*. Berlin: Springer.