SIOP-LEGIS: THESAURUS FOR SELECTION AND MANAGEMENT OF BRAZILIAN TREASURY DOMAIN

Stainam Brandão[1], Tiago Silva[1], Sergio Rodrigues[1], Luis Araujo[2], Daniel Silva[2] and Jano Souza[1]
[1] *COPPE/UFRJ, Graduate School of Engineeringy, Rio de Janeiro, Brazil*
[2] *Brazilian Department of the Treasury, Brasilia, Brazil*

Keywords: Thesaurus evolution, e-Government, Knowledge representation.

Abstract: This work presents the approach used to select documents from the Brazilian Official Gazette for Federal Treasury domain and manage the explicit knowledge through a thesaurus according to ISO Standard 2788. The goal of the SIOP-LEGIS project joint with Department of Treasury is to facilitate search and retrieval of documents semantic related through hierarquic terms defined in the Thesaurus. The paper also presents a brief case study to demonstrate how the ontology domain evolution ocurrs in accordance with the legislation, case law and administrative acts in constant changes.

1 INTRODUCTION

The Secretary of Federal Treasury, during the work, needs information to suport decision making and justify their actions legally. Likewise, access to legislation, jurisprudence and administrative actions is required from the Brazilian Official Gazette. However, read and follow the jurisprudence takes time and dedication. SIOP-LEGIS is a Knowledge Organizational System that mines Brazilian Official Gazette daily generating metadata for each document related to Federal Treasury domain.

Nowadays, SIOP-LEGIS repository has a collection of more than 80.000 indexed documents since 2008 and that number increases daily. However, only documents relevant to the domain are selected for further validation of experts.

Selection of Documents

Initially, domain experts have developed a thesaurus for the Federal Treasury with the main terms of Federal Budget area, whereas the system automatically selects the documents that would be part of through the use of thesaurus. After, the expert validates all of the selected documents, assigning one of three statuses:

- Validated: document selected automatically and validated by expert;
- Discarded: document automatically selected and discarded by an expert, not being a document of interest to Federal Budget area;
- Not selected: document not being selected (either automatically or manually).

Validation of Documents

There is a experts team in Treasury Department to perform the daily reading of the Federal Official Gazette and all decisions of the Supreme Court. This team also validate the documents selected and read the documents not selected for the evaluation of the methodology presented in this paper.

After validation, the document metadata created automatically for each validated document will be used as input for domain knowledge representation. At this point we present our approach through actions for knowledge representation and evolution, described in the next section.

2 REPRESENTATION AND EVOLUTION ACTIONS

Since the manual thesaurus construction requires a huge human effort, the acquisition of which is considered a bottleneck for the Knowledge Management. In this paper, domain thesaurus evolution involves enrichment through actions on the vocabulary, which is representing the explicit knowledge of Brazilian Official Gazettes to semantically index them.

This section presents a dynamic and adaptive representation through actions and rules that evolves

Brandão S., Silva T., Rodrigues S., Araujo L., Silva D. and Souza J..
SIOP-LEGIS: THESAURUS FOR SELECTION AND MANAGEMENT OF BRAZILIAN TREASURY DOMAIN.
DOI: 10.5220/0003660301950200
In *Proceedings of the International Conference on Knowledge Management and Information Sharing* (KMIS-2011), pages 195-200
ISBN: 978-989-8425-81-2

domains thesaurus identifying events those demands new configuration.

The indexing of documents in the domain thesaurus uses an approach to document adherence to each concept, through occurrence and proximity of terms. This work monitors events related to the quality metrics of domain ontologies and the changes occurring on the documents collection (*Corpus*) over time.

Our proposal handles domain thesaurus and these only treat the documents belonging to the same domain. There is no intention to identify the domain of a document and extend the thesaurus to another domain.

The next sub-sections present each action.

2.1 LDA / LSA

This action uses the topic modelling approach through the *Latent Dirichlet Allocation* - LDA (X. Wei e W. B. Croft, 2006) to identify topics for *Corpus* representation and after, applies the Latent Semantic Analysis - LSA (S. Cederberg e D. Widdows, 2003) approach to identify the terms correlated with higher values for latent semantic with the topics. These Topics and correlated terms are identified as candidates for the thesaurus concepts.

Applicability

The representation of words and sentences meaning captured by the LSA approach has been able to simulate the recognition of vocabulary and categorizing words, which make possible to approximate the human judgments of similarity between words, an aspect that is highly valued in the study of speech processing.

Model

Usually, topics are small and mostly consist of a single word (Watters D., 2009). Each topic represents ideas of origin, purpose or description of the documents content. In addition, the approach is applied on the 'term-document" matrix representation and applies the dimension reduction technique based on Singular Value Decomposition (SVD). In this moment, the documents and words are mapped into a representation of the latent semantic space, which is based on topics rather than each word individually and therefore the space of representation is much smaller than the original.

This control is temporal and every change in the documents collection leads to recalculated the topics and correlated terms to identify terms that have been outstanding, emerging and deprecating, according figure 1.

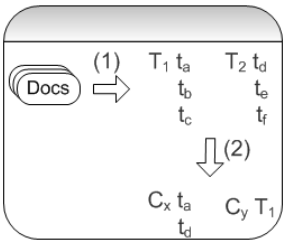


Figure 1: Mapping candidates terms: (1) mapping between topics (T) and correlated terms (t), (2) mapping between the thesaurus concepts © and topics (T) / correlated terms (t).

Example

After ran the crawler, register and pre-processing only the contents of the site of Brazilian Social Security (*http://www.previdenciasocial.gov.br/conteudoDinamico.php?id=690*), the proposed approach extracts and presents the topics and their respective correlated terms with the highest values of latent semantic for the experts. These terms correlated not only try to explain the meaning of the topic within the text but also serve as candidates to the concepts of domain thesaurus.

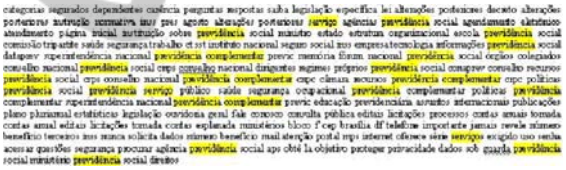


Figure 2: One of the Text pre-processing used for extracting topics.

Corpus - Collection of Terms Extracted

TOPIC	TFIDF	CORRELATED TERMS	
		TERM	PROB
COMPORTAMENTO	0.301	previdência	1.003
		acidentado acidente	1.002
		acidentado acidente	1.002
		cálculo	1.001
		versão	1.001
		benefícios benefício	1.001
FLUTUAÇÕES	0.301	previdência	1.003
		acidentado acidente	1.002
		acidentado acidente	1.002
		cálculo	1.001
		versão	1.001
		benefícios benefício	1.001

Figure 3: Two Topics with their correlated terms in Portuguese language: ‘Comportamento’ and ‘Flutuações’.

2.2 Concept Fragmentation

This event occurs when the thesaurus concepts have excess pointers to the documents indicating the need

for concept specialization with the identification of subclass in the collection of documents being referenced.

This overhead shows the importance of the concept for the domain and the possibility of new related concepts in the documents.

Applicability

The evolution must be based on clear rules to avoid inconsistency. And the Fragmentation acts on a concept considered important for the domain ontology due to the large number of documents indexed, which allows the identification of subclasses in documents.

This action has two strategies:

· Complete: if all subclasses cover all documents indexed, then the concept will be removed;

· Incomplete: If the subclasses do not cover all documents referenced, then the concept is retained in the domain ontology.

Model

The Fragmentation action uses guard expressions associated with the following metrics: Importance (Quantifies instances that belong to a concept in a sub-tree) and Class Richness (distribution of instances between concepts).

When some of these metrics achieve a value that triggers at least one of the guard expressions, the Fragmentation action is performed:

1. Identify the concept that triggered the guard expression;
2. Select the Terms Correlated with greater similarity with the concept (identified in step 1) in the documents indexed;
3. Analysis the completeness of the Correlated of Terms:

a. If the subclasses are complete, then the concept identified in step 1 is removed;

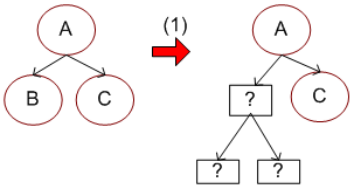


Figure 4: Strategy: Complete.

b. If the subclasses are incomplete, so the concept identified in step 1 is kept in the model.

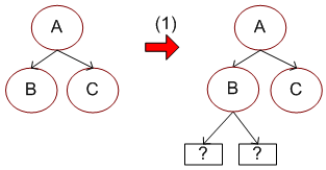


Figure 5: Strategy: Incomplete.

4. Redistribution of the documents indexed among the concepts involved. Using the approach to document adherence to each concept, through occurrence and proximity of terms.

Example

The figure 6 shows a hypothetical concept 'Pensão' (‘pension’, in English) which indexes a collection of documents and has no subclasses. Whilst figure 7, shows the identification of ten subclasses that represent faithfully the documents collection previously referenced by the term 'Pensão' deleted.

Figure 6: Concept ‘Pensão’ has no subclass.

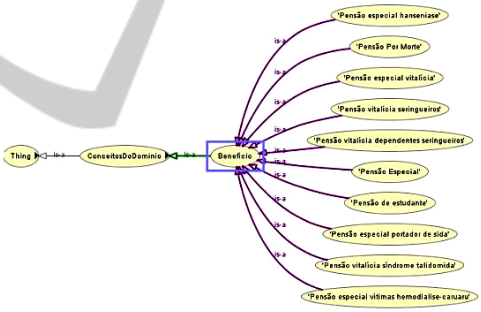


Figure 7: Fragmentation action: ten subclasses identified and concept ‘Pensão’ deleted.

2.3 Abstraction

This action transforms the domain ontology in a faithful representation of the Corpus. However, the knowledge already represented cannot be lost, because each application uses the knowledge that supports your desired area. So the idea is to use semantic relativity in the model in an accordance of the necessity of the application. Therefore, a versioning control is used.

Applicability

The applicability is related to ability to provide thesaurus abstractions (views) for applications that operate in different areas of the same domains represented by the ontology.

Model

The model uses the Corpus to identify the thesaurus concepts missing as well as highlights those has its co-occurrence and frequency decreased overtime.
For this, future approaches will be used TF-IDF and LDA / LSA.

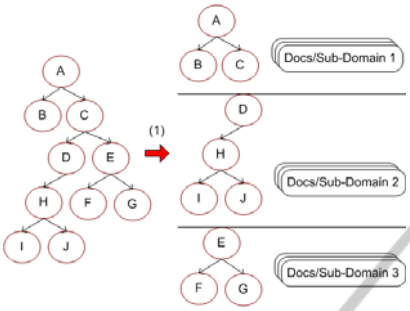


Figure 8: Abstraction per sub-domain.

Example

Within the Federal Treasury domain, a thesaurus was developed initially for the Federal Budget area, due to a system for registering, searching and retrieving documents from the Brazilian Official Gazette. However, with the initial success of the proposal, two other distinct areas of Federal Treasury domain have expressed interest in using the system: Human Resources Secretary (SRH) and Shareholders' Union Secretary (SHF). Thus, this action provides conditions from a single thesaurus providing three abstractions (views) of the domain to different areas that may have common concepts as distinct concepts.

3 SIOP-LEGIS

The Knowledge Organizational System (SIOP-LEGIS) is a project that aims to provide knowledge management for legislative domain through changeable representation, which deal trends in the law. Its architecture is presented in Figure 9.

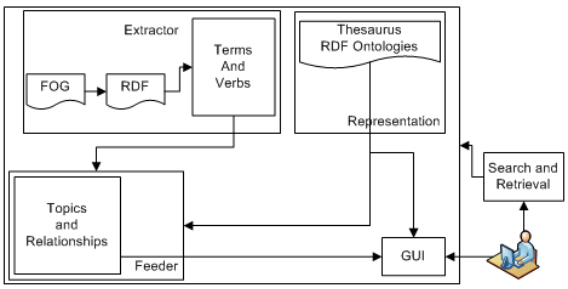


Figure 9: Initial Architecture.

Natural Language Processing

Each known document from Brazilian Official Gazettes has a type: *Order*, *Law*, *Internal Program*, *etc*. Since the relationships initially defined between them are: *revoke*, *alter*, *establish*, *regulate and deploy and create*. All of those are modelled in an RDF ontology for not to be considered during the natural language processing. Likewise stopwords and list of known entities that contains authorities (people and organizations), technical terms, dates and values that are not of interest for representation in the thesaurus

4 CASE STUDY

This section presents a brief case study to demonstrate how the architecture works. The example used in this work is adapted from Secretary of Federal Treasury responsible for control and oversight of federal spending in accordance with the legislation, case law and administrative acts.
For this, the SIOP-LEGIS system was used for semi-automatic selection and indexing of documents from Brazilian Official Gazettes that deals with the Federal Treasury domain.

Ministério da Integração Nacional

GABINETE DO MINISTRO

PORTARIA Nº 571, DE 12 DE JULHO DE 2010

•
•
•

Art. 4o Na elaboração da proposta de aplicação dos recursos do FCO, a ser encaminhada pelo Banco do Brasil ao Ministério da Integração Nacional (MI), até 30 de setembro de 2010, serão observadas as seguintes orientações de caráter geral:

I - deverá ser concedido tratamento diferenciado e favorecido, no que diz respeito ao percentual de limite de financiamento, aos projetos de mini e pequenos produtores rurais e de micro e pequenas empresas beneficiárias do FCO, bem como aos empreendimentos que se localizem nos espaços prioritários da PNDR;

II - a proposta de aplicação dos recursos do FCO para o exercício de 2011 deverá ser formulada pelo Banco do Brasil, em articulação com a Secretaria de Desenvolvimento do Centro-Oeste (SCO) do Ministério da Integração Nacional (MI);

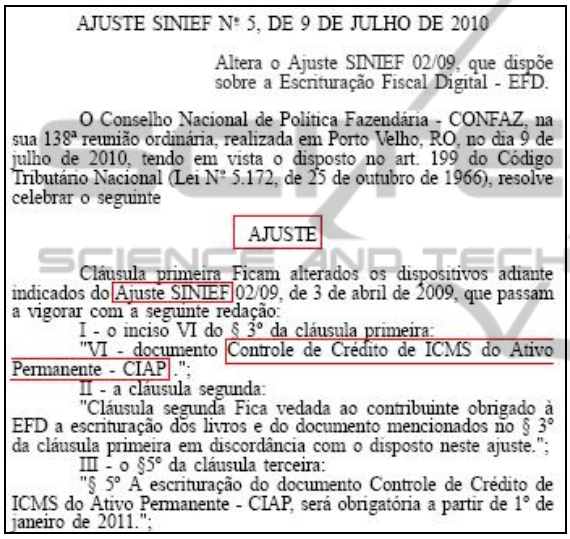
III - a proposta de aplicação dos recursos do FCO deverá apresentar quadro demonstrativo do orçamento previsto para o exercício de 2011, estimando a totalidade dos ingressos e das saídas de recursos previstos para o ano, especificando:

Figure 10: This caption shows de topics extracted from a validated document.

From the domain thesaurus of the Department of Treasury, the system automatically pre-selects the documents that would be part of this domain. After, experts validate the pre-selected documents that acquire one of three statuses:

(1) Validated: document selected automatically and validated by an expert;
(2) Discarded: document automatically selected and discarded by expert not being a document of interest;
(3) Not selected: document not being selected automatically.

Each validated document was indexed on thesaurus terms and used by the LDA / LSA action for identification of new terms and relations. Figure 10 and 11 show two documents validated with the terms identified by LDA / LSA for the domain thesaurus evolution.



AJUSTE SINIEF Nº 5, DE 9 DE JULHO DE 2010

Altera o Ajuste SINIEF 02/09, que dispõe sobre a Escrituração Fiscal Digital - EFD.

O Conselho Nacional de Política Fazendária - CONFAZ, na sua 138ª reunião ordinária, realizada em Porto Velho, RO, no dia 9 de julho de 2010, tendo em vista o disposto no art. 199 do Código Tributário Nacional (Lei Nº 5.172, de 25 de outubro de 1966), resolve celebrar o seguinte

AJUSTE

Cláusula primeira Ficam alterados os dispositivos adiante indicados do Ajuste SINIEF 02/09, de 3 de abril de 2009, que passam a vigorar com a seguinte redação:

I - o inciso VI do § 3º da cláusula primeira:

"VI - documento Controle de Crédito de ICMS do Ativo Permanente - CIAP.";

II - a cláusula segunda:

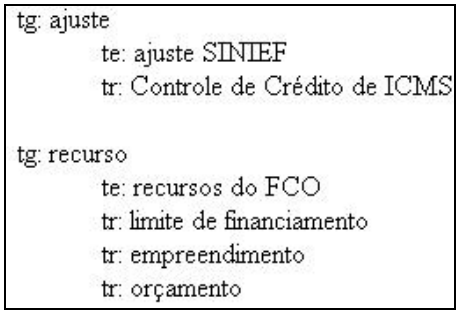
"Cláusula segunda Fica vedada ao contribuinte obrigado à EFD a escrituração dos livros e do documento mencionados no § 3º da cláusula primeira em discordância com o disposto neste ajuste.";

III - o §5º da cláusula terceira:

"§ 5º A escrituração do documento Controle de Crédito de ICMS do Ativo Permanente - CIAP, será obrigatória a partir de 1º de janeiro de 2011.";

Figure 11: This caption shows another topics extracted from a validated document that will enrich the thesaurus.

The methodology generates a new version of thesaurus with the inclusion of the following knowledge represented.



tg: ajuste
te: ajuste SINIEF
tr: Controle de Crédito de ICMS

tg: recurso
te: recursos do FCO
tr: limite de financiamento
tr: empreendimento
tr: orçamento

Figure 12: This caption show de thesaurus evolution from the two documents previously captured, according ISO 2788. *TG*: concepts with wider connotation; TE: concepts with more specific connotation; TR: concepts with more specific connotation.

Currently, the domain thesaurus is used as artefact on the recommendation mechanism for search and retrieval in the same system.

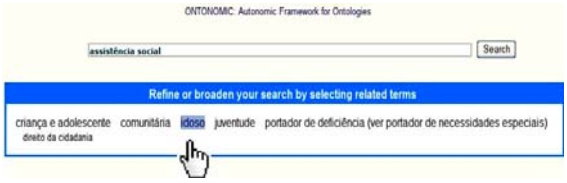


Figure 13: Recommendation mechanism for search.

5 CONCLUSIONS AND FUTURE WORKS

This paper presents a methodology applied by Secretary of Federal Budget for knowledge representation and management of Brazilian Official Gazette for Federal Treasury domain. A domain thesaurus is used to represent a changeable domain with a daily inclusion of new Gazette with legislation, jurisprudence and administrative actions.

One future work is the development of a Context-Sensitive Search based on the content documents analysed, with identification of resources and relationship previously specified on domain and represented on RDF ontology.

Another future work is allow navigation through documents by these RDF resources, deeming the correlation of documents already defined with the Vector Model (TF-IDF) approach from (Lucene, 2009) to calculate the similarity between the documents.

A future target is the integration with Formal Ontology to establish formal meanings for domain vocabulary allowing axiomatization and integration of domain ontologies from different sources. We notice that domain ontologies consist of specialized terminology and a particular vision of reality. But the meaning remains dependent on the context. The use of Formal Ontology to integrate the domain ontology can promote the reuse, integration and management through the construction of ontologies from different countries about the same domain.

ACKNOWLEDGEMENTS

We thank CAPES, CNPq and Department of Treasury for supporting this work.

REFERENCES

A. Maedche, B. Motik, L. Stojanovic, R. Studer, e R. Volz, “An infrastructure for searching, reusing and evolving distributed ontologies,” WWW, 2003, pp. 439–448.

B. Omelayenko, “Learning of Ontologies for the Web: the Analysis of Existent Approaches,” *Proceedings of the International Workshop on Web Dynamics, held in conj. with the 8th International Conference on Database Theory* (ICDT’01), London, UK, 2001.

Clips, “http://clipsrules.sourceforge.net/” Acessed in august 22th 2009.

G. Guizzardi, “Ontological foundations for structural conceptual models,” 2005.

H. Alani e C. Brewster, “Ontology ranking based on the analysis of concept structures,” *Proceedings of the 3rd international conference on Knowledge capture*, Banff, Alberta, Canada: ACM, 2005, pp. 51-58.

Jaccard, “http://sourceforge.net/projects/simmetrics/” Acessed in august 22th 2009.

Jess, “http://www.jessrules.com/” Acessed in august 22th 2009.

K. Beyer, J. Goldstein, R. Ramakrishnan, e U. Shaft, “When Is "Nearest Neighbor" Meaningful?,” *In Int. Conf. on Database Theory*, 1999, pp. 217--235.

M. W. Geert, G., “The Ontological Foundation of REA Enterprise Information Systems,” Ago. 2000.

N. Guarino, “Formal Ontology and Information Systems,” 1998.

M. Shaw e D. Garlan, Software Architecture: Perspectives on an Emerging Discipline, Prentice Hall, 1996.

N. Russell, A. ter Hofstede, D. Edmond, e W. van der Aalst, “Workflow Data Patterns: Identification, Representation and Tool Support,” Conceptual Modeling – ER 2005, 2005, pp. 353-368.

N. Russell, Arthur, W. van der Aalst, e N. Mulyar, Workflow Control-Flow Patterns: A Revised View, 2006.

S. A. Macskassy, A. Banerjee, B. D. Davison, e H. Hirsh, “Human Performance on Clustering Web Pages: A Preliminary Study,” KDD, 1998, pp. 264-268.

T. Berners-Lee, J. Hendler, e O. Lassila, “The Semantic Web A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities,” 2001.

T. R. Gruber, “A translation approach to portable ontology specifications,” Knowl. Acquis., vol. 5, 1993, pp. 199

Gomez-Perez, A., Knowledge sharing and reuse, 1998. In J. Liebowitz, ed., The Handbook of Applied Expert Systems, CRC Press.

Lucene, “http://lucene.apache.org/” Acessed in november 12th 2009.

Watters D., “Meaningful Clouds: Towards a novel interface for document visualization”. Disponível em: <http://danwatters.com/documents/CloudMine_dwatters.pdf>. Acesso em 09/09/2010, 2009.

S. Cederberg e D. Widdows, “Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction”, in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* - Edmonton, Canada, 2003, pp. 111-118.

X. Wei e W. B. Croft, “LDA-based document models for ad-hoc retrieval”, in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2006, p. 178–185.