Pontifícia Universidade Católica do Rio Grande do Sul
Programa de Pós-Graduação em Ciência da Computação
Grupo de Pesquisa em PLN

III Workshop on MSc dissertation and
PhD thesis in Artificial Intelligence - WTDIA'2006

Uma Medida de Similaridade Semântica entre Ontologias em Língua Portuguesa

Juliano Baldez de Freitas
jfreitas@inf.pucrs.br

Vera Lúcia Strube de Lima
vera@inf.pucrs.br

Agenda

- **Ontologias**
- **Mapeamento entre ontologias**
- **Proposta de Trabalho**
  - Mapeamento Taxonômico (Maedche e Staab)
  - Medida de Similaridade Semântica (SiSe)
  - Módulo SiSe
- **Avaliação**
- **Considerações**
- **Referências**

Ontologias

- termo "ontologia" na Ciência da Computação teve origem na comunidade de Inteligência Artificial (IA);

- [Fensel 2002], [Holsapple e Joshi 2002], [Chandrasekaran, Josephson e Benjamins 1999], chegam a um consenso sobre ontologias:

  "*uma ontologia identifica classes - cada uma caracterizada por propriedades que todos os elementos desta classe compartilham -e as organiza hierarquicamente. Isto também inclui importantes relações entre classes e elementos, em um domínio de conhecimento específico*."

- em [Gruber:95], [Guarino:96], o autor define uma ontologia como: "*uma especificação explícita e formal de uma conceitualização compartilhada*"

Mapeamento entre Ontologias

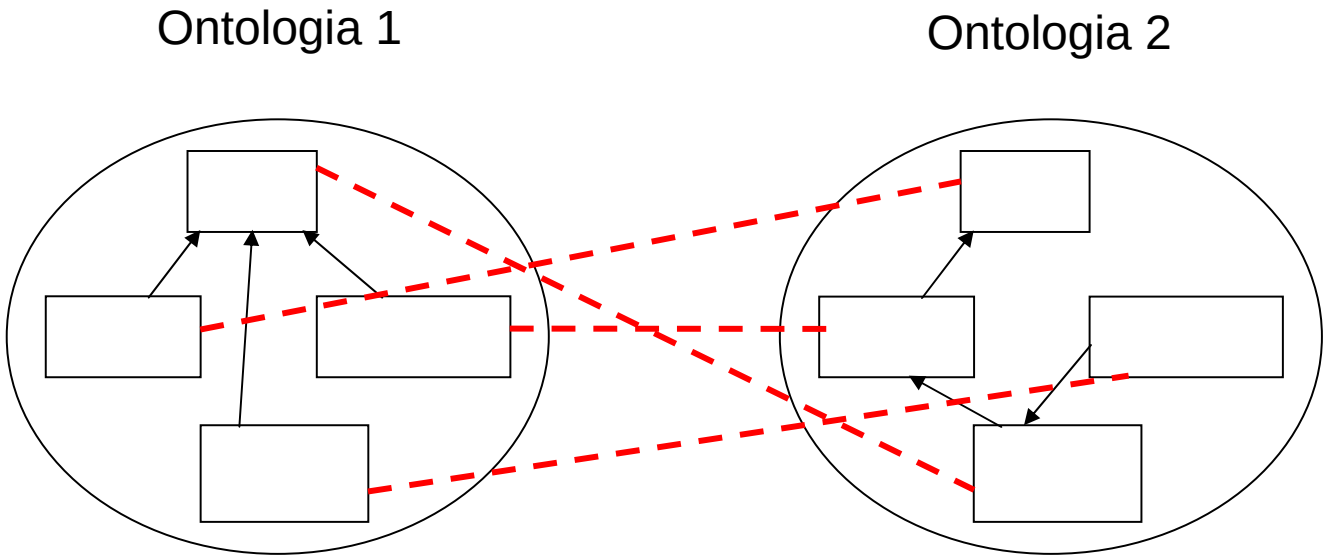
- **O que é:**

  O mapeamento entre duas ou mais ontologias associa conceitos e relações equivalentes, de diferentes origens, uns com os outros, de acordo com relações de similaridade.

- **Para que serve:**
  - 1) o uso e reuso de ontologias;
  - 2) expansão e combinação das mesmas, com o intuito de aumentar a informação e conhecimento em diferentes domínios que são integrados para suportar nova comunicação e uso.

- é uma operação crítica em muitos domínios: Web Semântica, *schemas* XML, banco de dados, ontologias;

Mapeamento entre Ontologias



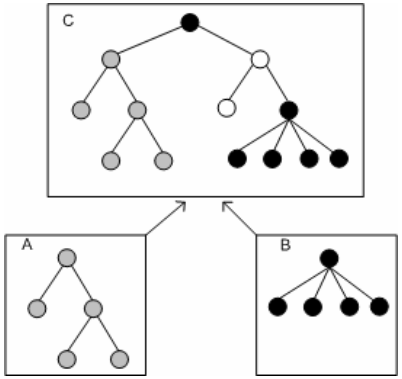
Mapeamento entre Ontologias

- **Características analisadas [Noy 2004]:**
  - nomes dos conceitos e descrições em linguagem natural;
  - hierarquia das classes (relacionamentos de subclasses e superclasses);
  - definições de propriedades (domínio, abrangência, restrições);
  - instâncias das classes;
  - descrições das classes.

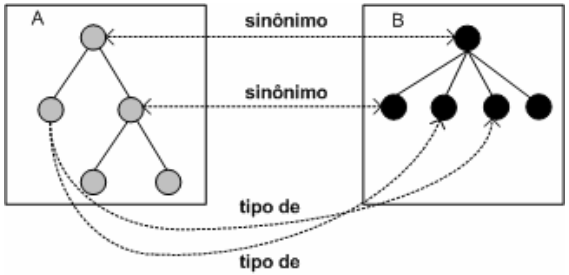
- **Problemas [Noy 2004] [Maedche e Staab 2002]:**
  - conceitualizações em linguagem natural;
  - hierarquias distintas

Mapeamento entre Ontologias

- **união:**

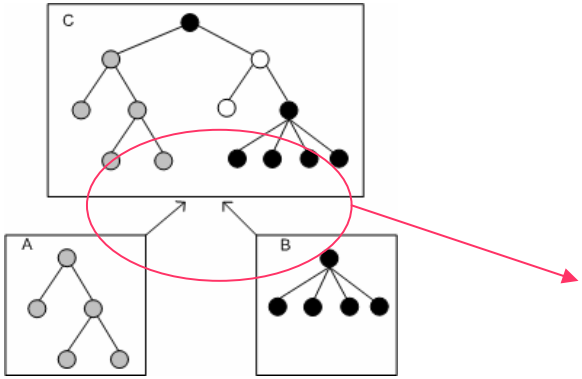


- **alinhamento:**



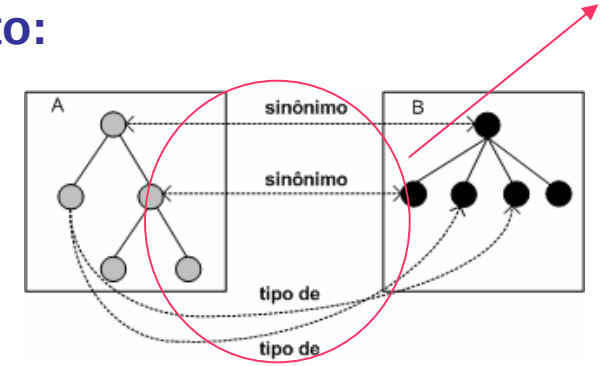
Mapeamento entre Ontologias

- **união:**



Medidas de Similaridade

- **alinhamento:**



Proposta de Trabalho

- **Mapeamento Taxonômico (MT) [Maedche e Staab 2002]:**
  - compara a estrutura hierárquica de duas ontologias;
  - Utiliza na comparação as relações de hierarquias dos nodos;
  - “Semantic cotopy”:

$$SC(C_i, H) := \{C_j \in A | H(C_i, C_j) \vee H(C_j, C_i) \vee C_j = C_i)\}$$

  - onde H é a taxonomia, e H(Ci,Cj) significa que Ci é um subconceito de Cj , e A é um conjunto de conceitos da ontologia;
  - utiliza a Medida de Jaccard para conjuntos;

Proposta de Trabalho

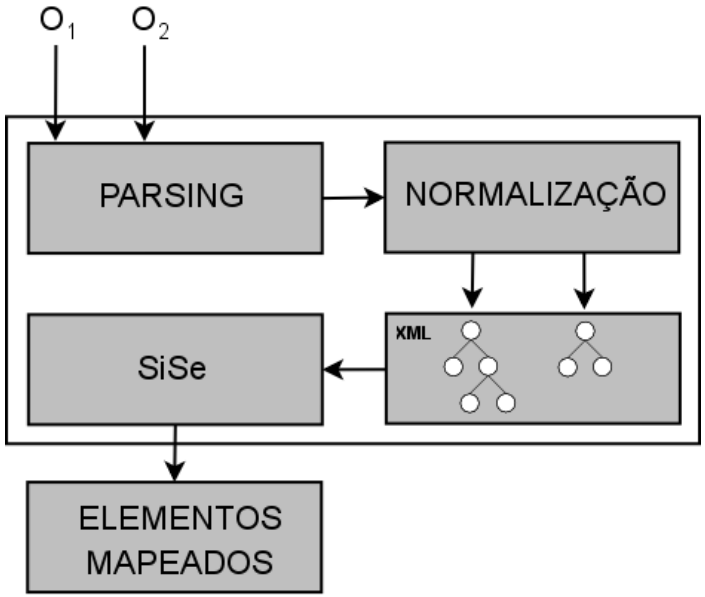
- **Medida de Similaridade Semântica (SiSe):**
  - adaptação da medida de [Maedche e Staab 2002] Mapeamento Taxonômico;
  - adaptação do conceito de "Semantic Cotopy" (SC):

$$SC(C_i, H) := \{\triangle C_j \in A | H(C_i, C_j) \vee H(C_j, C_i) \vee C_j = C_i)\}$$

  - fases de execução da medida: *parsing*, normalização, algoritmo de mapeamento, indicação dos mapeamentos;
  - utiliza recursos lingüísticos: *stemming, heurísticas;*
  - *coeficiente de similaridade entre 0 e 1,* onde 1 representa uma combinação perfeita;
  - criação de heurísticas para aumentar o coeficiente de similaridade;

Proposta de Trabalho

- **Estratégia da medida SiSe:**



Proposta de Trabalho

- **PARSING:**
  - analisa a linguagem utilizada na representação da ontologia (TML, RDF, OWL, DAML);
  - extrai relações de hierarquia:
    - Hiponímia: subconceitos
    - Hiperonímia: superconceitos
  - exemplo:

```
<?xml version="1.0" ?>
<THESAURUS tipo="VCBS">
  <T term="DIREITO ELEITORAL">
    <BT term="DIREITO CONSTITUCIONAL"/>
    <NT term="CAMPANHA ELEITORAL"/>
    <NT term="ELEICAO"/>
    <NT term="PARTIDO POLITICO"/>
    <NT term="SISTEMA ELEITORAL"/>
    <NT term="VOTO"/>
  </T>
</THESAURUS>
```

Relações de Hierarquia:

- **T: Term**
- **BT: Broader Term**
- **NT: Narrower Term**

Proposta de Trabalho

- **NORMALIZAÇÃO:**
  - reproduz a ontologia em formato XML;
  - permite a representação em forma de árvore;
  - abstrai a sintaxe da linguagem da ontologia
  - comparação da similaridade: hierarquia x hierarquia;
  - exemplo:

```
<?xml version="1.0" ?>
<THESAURUS tipo="VCBS">
  <T term="DIREITO ELEITORAL">
    <BT term="DIREITO CONSTITUCIONAL"/>
    <NT term="CAMPANHA ELEITORAL"/>
    <NT term="ELEICAO"/>
    <NT term="PARTIDO POLITICO"/>
    <NT term="SISTEMA ELEITORAL"/>
    <NT term="VOTO"/>
  </T>
</THESAURUS>
```

```
<?xml version="1.0" ?>
<ontologia>
  <classe> DIREITO ELEITORAL
    <subclasse> DIREITO CONSTITUCIONAL
      <subclasse>CAMPANHA ELEITORAL </subclasse>
      <subclasse>ELEICAO</subclasse>
      <subclasse>PARTIDO POLITICO</subclasse>
      <subclasse>SISTEMA ELEITORAL</subclasse>
      <subclasse>VOTO</subclasse>
    </subclasse>
  </classe>
</ontologia>
```

Proposta de Trabalho

- **SiSe**

Ontologia 1 (O_1)

- direito constitucional
  - direito eleitoral
    - campanha eleitoral
    - eleição
    - partido político
    - sistema eleitoral
    - voto

Ontologia 2 (O_2)

- direito
  - direito eleitoral
    - crime eleitoral
    - domicílio eleitoral
    - eleições
    - justiça eleitoral
    - partidos políticos
    - sistema distrital
    - voto

Proposta de Trabalho

- **exemplo**

Ontologia 1 (O_1)

- direito constitucional
  - direito eleitoral
    - campanha eleitoral
    - eleição
    - partido político
    - sistema eleitoral
    - voto

superconceitos

stemming:

eleição → ele

direito eleitoral → direitEleitor

direito constitucional → direitConstituc

Semantic Cotopy:

SC(eleição, O_1) =

{ele, direitEleitor, direitConstituc}

Proposta de Trabalho

- **exemplo**

Ontologia 2 (O_2)

superconceitos

- direito
  - direito eleitoral
    - crime eleitoral
    - domicílio eleitoral
    - eleições
    - justiça eleitoral
    - partidos políticos
    - sistema distrital
    - voto

stemming:

eleições → ele

direito eleitoral → direitEleitor

direito → direit

Semantic Cotopy:

SC(eleição, O_2) =

{ele, direitEleitor, direit}

Proposta de Trabalho

- **exemplo:**

| SC(eleição, O1) ∩ SC(eleições, O2) |

| {ele, direitEleitor, direitConstituc} ∩ {ele, direitEleitor, direit} |

→ | {ele, direitEleitor} | = 2

| SC(eleição, O1) ∪ SC(eleições, O2) |

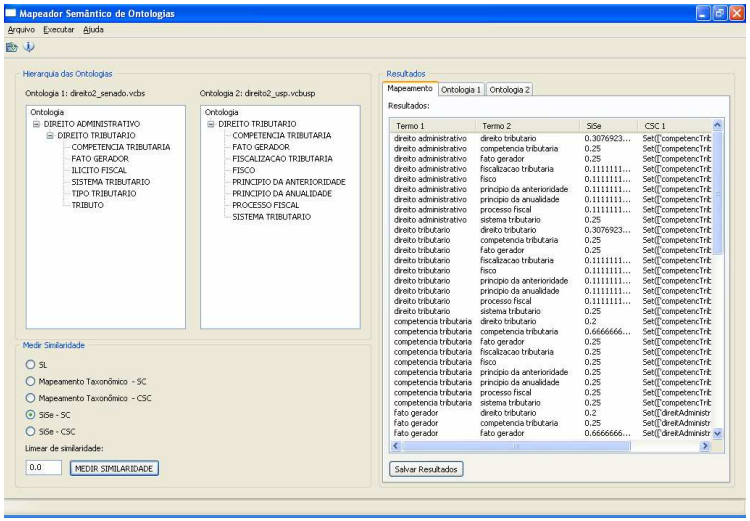
| {ele, direitEleitor, direitConstituc} ∪ {ele, direitEleitor, direit} |

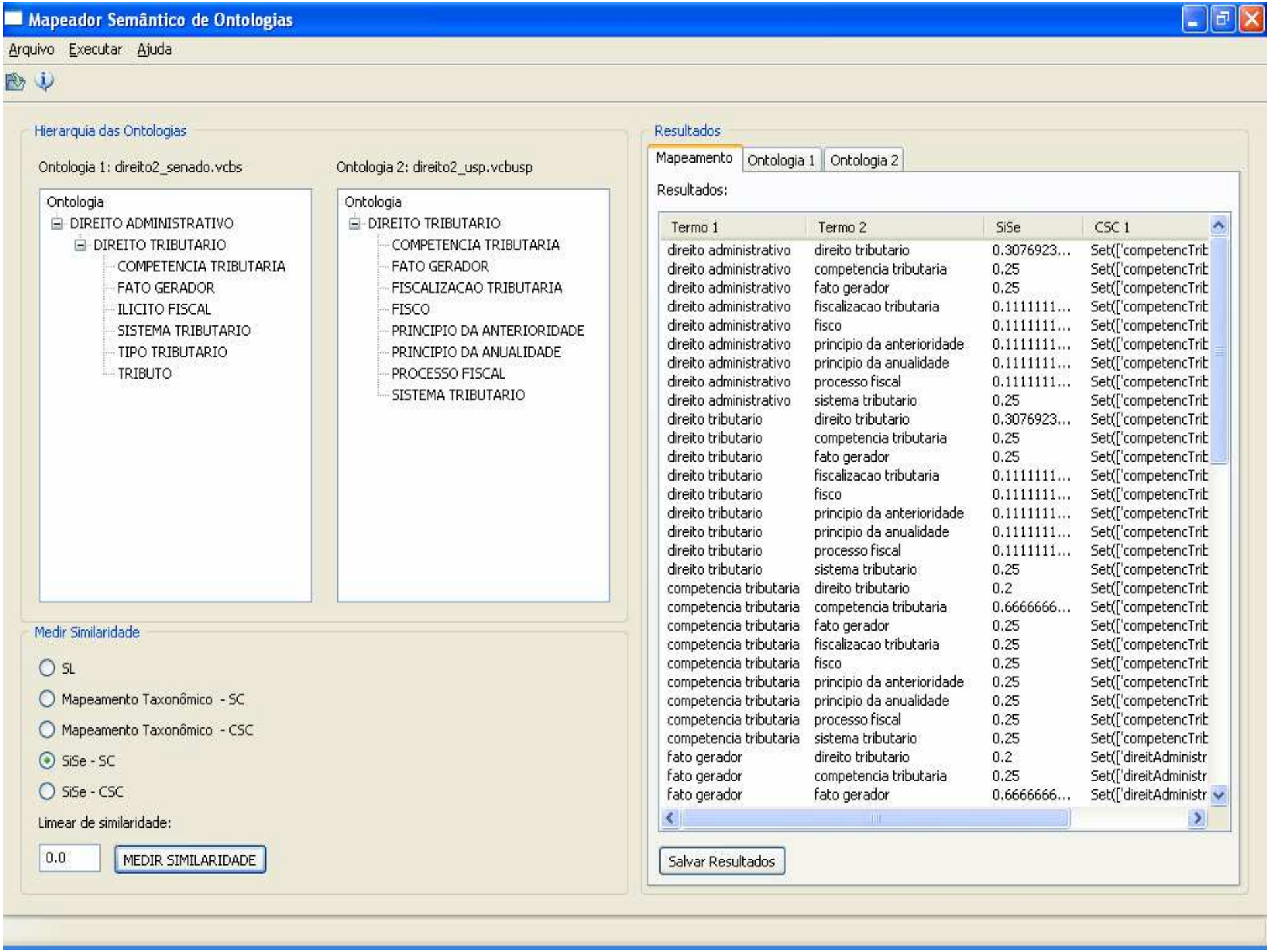
→ | {ele, direitEleitor, direitConstituc, direit} | = 4

2/4 = 0.5 ∈ [0,1]

Proposta de Trabalho

- **Módulo SiSe:**
  - módulo Python para medir a similaridade entre ontologias;
  - implementa a estrátegia descrita anteriormente;
  - interface gráfica para análise dos resultados:




Mapeador Semântico de Ontologias
Arquivo Executar Ajuda
Hierarquia das Ontologias
Ontologia 1: direito2_senado.vcbs
Ontologia
DIREITO ADMINISTRATIVO
DIREITO TRIBUTARIO
COMPETENCIA TRIBUTARIA
FATO GERADOR
ILICITO FISCAL
SISTEMA TRIBUTARIO
TIPO TRIBUTARIO
TRIBUTO
Ontologia 2: direito2_usp.vcbusp
Ontologia
DIREITO TRIBUTARIO
COMPETENCIA TRIBUTARIA
FATO GERADOR
FISCALIZACAO TRIBUTARIA
FISCO
PRINCIPIO DA ANTERIORIDADE
PRINCIPIO DA ANUALIDADE
PROCESSO FISCAL
SISTEMA TRIBUTARIO
Medir Similaridade
SL
Mapeamento Taxonômico - SC
Mapeamento Taxonômico - CSC
SiSe - SC
SiSe - CSC
Limear de similaridade:
0.0
MEDIR SIMILARIDADE
Resultados
Mapeamento
Ontologia 1
Ontologia 2
Resultados:
Termo 1 Termo 2 SiSe CSC 1
direito administrativo direito tributario 0.3076923... Set(['competencTrit
direito administrativo competencia tributaria 0.25 Set(['competencTrit
direito administrativo fato gerador 0.25 Set(['competencTrit
direito administrativo fiscalizacao tributaria 0.1111111... Set(['competencTrit
direito administrativo fisco 0.1111111... Set(['competencTrit
direito administrativo principio da anterioridade 0.1111111... Set(['competencTrit
direito administrativo principio da anualidade 0.1111111... Set(['competencTrit
direito administrativo processo fiscal 0.1111111... Set(['competencTrit
direito administrativo sistema tributario 0.25 Set(['competencTrit
direito tributario direito tributario 0.3076923... Set(['competencTrit
direito tributario competencia tributaria 0.25 Set(['competencTrit
direito tributario fato gerador 0.25 Set(['competencTrit
direito tributario fiscalizacao tributaria 0.1111111... Set(['competencTrit
direito tributario fisco 0.1111111... Set(['competencTrit
direito tributario principio da anterioridade 0.1111111... Set(['competencTrit
direito tributario principio da anualidade 0.1111111... Set(['competencTrit
direito tributario processo fiscal 0.1111111... Set(['competencTrit
direito tributario sistema tributario 0.25 Set(['competencTrit
competencia tributaria direito tributario 0.2 Set(['competencTrit
competencia tributaria competencia tributaria 0.6666666... Set(['competencTrit
competencia tributaria fato gerador 0.25 Set(['competencTrit
competencia tributaria fiscalizacao tributaria 0.25 Set(['competencTrit
competencia tributaria fisco 0.25 Set(['competencTrit
competencia tributaria principio da anterioridade 0.25 Set(['competencTrit
competencia tributaria principio da anualidade 0.25 Set(['competencTrit
competencia tributaria processo fiscal 0.25 Set(['competencTrit
competencia tributaria sistema tributario 0.25 Set(['competencTrit
fato gerador direito tributario 0.2 Set(['direitAdministr
fato gerador competencia tributaria 0.25 Set(['direitAdministr
fato gerador fato gerador 0.6666666... Set(['direitAdministr
Salvar Resultados

Avaliação

- ***Golden Mapping*:**
  - avaliação humana feita antes da aplicação da medida SiSe;
  - criação de um "Golden Mapping" (ou mapeamento dourado) para avaliação;
    - humano pode encontrar mapeamentos que a medida não encontra;
    - abordagem menos tendenciosa;
    - deixar disponível para a comunidade este recurso para futuras avaliações de medidas de similaridade que venham a ser desenvolvidas;
  - ontologias do domínio do Direito;
  - 5 pares de ontologias analisadas por três humanos
    - um lingüísta, um cientista da computação, um bacharel em direito;
  - comparação SiSe x Humanos.

Considerações

- SiSe é uma medida de similaridade semântica entre ontologias voltadas ao português;
- a medida SiSe refina o coeficiente de similaridade da abordagem de Maedche e Staab;
- é necessária a criação de heurísticas que aumentem o coeficiente de similaridade em alguns casos;
- uso de dicionários de sinônimos para melhorar a medida;
- avaliação em ontologias de outros domínios;
- dificuldades:
  - falta de repositório de ontologias em língua portuguesa.

Referências mais importantes

- [Fensel 2002]FENSEL, D. Ontology-based knowledge management. IEEE Computer, v. 35, n. 11, p. 5659, November 2002.
- [Gruber 1995]GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing. In: International Journal of Human-Computer Studies, v. 43, n. 5/6, p. 907-928, 1995.
- [Guarino 1996]GUARINO, N. Understanding, building, and using ontologies. a commentary to using explicit ontologies in kbs development. In: Proceedings of the 10th KnowledgeAquisition for Knowledge-Based Systems Workshop, n. Ban, p. Canada, 1996.
- [Holsapple e Joshi 2002]HOLSAPPLE, C. W.; JOSHI, K. D. A collaborative approach to ontology design. Communications of the ACM, v. 45, n. 2, p. 42-47, February 2002.
- [Maedche e Staab 2002]MAEDCHE, A.; STAAB, S. Measuring similarity between ontologies. In: Proceedings of the European Conference on Knowledge Acquisition and Management (EKAW), 2002.
- [Noy 2004]NOY, N. F. Semantic integration: a survey of ontology-based approaches. SIGMOD Record, v. 33, n. 4, p. 6570, December 2004.

? Perguntas ?

- **Página do Projeto:**
  - http://www.inf.pucrs.br/~jfreitas
- **Contatos:**
  - jfreitas@inf.pucrs.br
  - vera@inf.pucrs.br
- **Bolsa:**
  - CDPe - Centro de Desenvolvimento e Pesquisa Dell-PUCRS