Prova de Qualificação de Doutoramento em Informática
Faculdade de Ciências da Universidade de Lisboa

Uma metodologia para construção de ontologias e integração de conhecimento geográfico

Marcirio Silveira Chaves

Orientadores
Mário J. Silva e Diana Santos

Motivação

- Necessidade de metodologia para construção de ontologias geográficas
- Conhecimento presente em textos não é suficientemente legível por máquina
- Ontologias desconexas na web semântica
- Ontologias carecem de ocorrências
- Pouco conhecimento sobre a terminologia geográfica utilizada em textos em português

Dificuldades na construção de ontologias geográficas

- Fontes de informação geográfica publicamente disponíveis são ricas em informação, mas **incompletas**
- A LN é **vaga** e **ambígua**, o que dificulta o processo de extração e representação de conhecimento
- Conceitos geográficos ricos em propriedades
- Conhecimento geográfico **incerto**, **incompleto**, **contraditório** e **variável com o tempo**

Estrutura da apresentação

- Conceitos
- Estado da arte
- Objetivos
  - Contribuições
- Trabalho realizado e em andamento
- Sistema de Extração e Integração de conhecimento geográfico – SEI-Geo
- Avaliação
- Plano, Cronograma e Marcos

Áreas envolvidas na tese

Conceitos

- Representação de conhecimento
  - Diferentes visões (geógrafos, informáticos, senso comum, etc.)
  - Ontologia
    - Representação estruturada e compartilhada do conhecimento geográfico
    - De autoridades e textos

Conceitos

- Extensão de ontologias
  - Aquisição de **novos** conceitos e relacionamentos a partir do texto e posteriormente acrescentados a uma ontologia existente.
  - 2 níveis:
    - nível conceitual: novos conceitos e relacionamentos detectados em texto
    - nível de termos: ocorrências dos conceitos adicionados à ontologia existente (Povoamento de ontologias)

Extração de Informação

- EI é o processo de obtenção de **dados quantificáveis desambiguados** a partir da LN, para servir a alguma necessidade de informação **precisa e pré-especificada** (Cunningham 06).
  - informação precisa e pré-especificada: conceitos, atributos e relacionamentos dentro de uma ontologia geográfica

Integração de Dados x Informação

- Integração de dados no domínio geográfico
  - Diferentes abordagens para conceitos e ocorrências
    - Conceitos -> medidas de similaridade lexical
    - Ocorrências -> combinação perfeita
- Integração de informação geográfica
  - Considera relacionamentos
  - Integração da hierarquia proveniente do texto com a existente na ontologia geográfica

Integração de Informação

O rio Douro (Duero, em castelhano) é um rio que nasce em Espanha, na província de Sória, nos picos da Serra de Urbião (Sierra de Urbión), a 2.080 metros de altitude e atravessa o norte de Portugal. A foz do Douro é junto à cidade do Porto. Tem 850 km de comprimento. Afluentes: Rio Paiva, Rio Sousa, Rio Tua.

- Conceitos: rio e serra
- Propriedades: altitude e comprimento
- Conceitos e propriedades já estão definidos na ontologia e o povoamento é feito conforme a presença das ocorrências no texto.

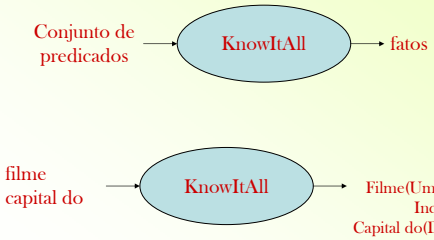
Integração de Informação

```
<gn:Geo_Feature rdf:ID="GEO_238">
  <gn:names>
    <gn:name="Porto" xml:lang="PT-PT" gn:att="P" gn:is="INE"/>
  </gn:names>
  <gn:geo_type_id rdf:resource="#CON"/>
  ...
</gn:Geo_Feature>
<gn:Geo_Feature rdf:ID="GEO_PHY_145">
  <gn:names>
    <rdf:Bag>
      <rdf:li gn:name="Douro" xml:lang="PT-PT" gn:att="P" gn:is="WIKI"/>
      <rdf:li gn:name="Duero" xml:lang="ES-ES" gn:att="A" gn:is="WIKI"/>
    </rdf:Bag>
  </gn:names>
  <gn:geo_type_id gn:is="WIKI" rdf:resource="#Rio"/>
  <rdfs:comment>Serra de Urbião - Spain</rdfs:comment>
  <gn:source_river gn:is="WIKI" rdf:resource="#GEO_PHY_120"/>
  <rdfs:comment>Porto - Portugal</rdfs:comment>
  <gn:outlet_river gn:is="INE" rdf:resource="#GEO_ADM_238"/>
  <gn:tributary gn:is="WIKI">
    <rdf:Bag>
      <rdf:li rdf:resource="#GEO_PHY_400"/>
      <rdf:li rdf:resource="#GEO_PHY_401"/>
    </rdf:Bag>
  </gn:tributary>
  <gn:length unit="km" gn:is="WIKI">850</gn:length>
```

Conhecimento existente

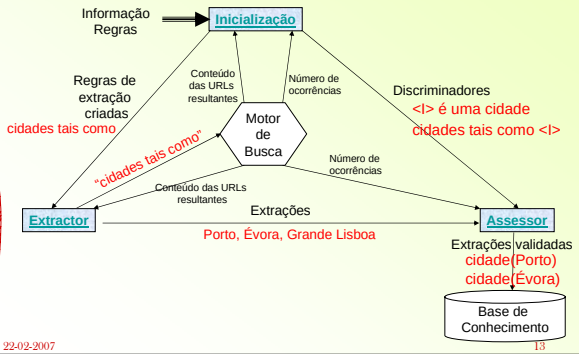
Estado da Arte

- KnowItAll



Estado da Arte

- KnowItAll



Estado da Arte

- **OntoSyphon**
  - Recebe um conjunto de conceitos de uma ontologia
  - Procura ocorrências desses conceitos em textos da web com o auxílio de padrões léxico-sintáticos (Hearst, 1992)
  - Usa BE (*Binding Engine*) para consultas
    - Ex.: "cidades tais como **<SN>**"
  - Atribui um grau de confiança às ocorrências extraídas baseado na freqüência de cada ocorrência nos textos
  - Domínios: Animal, Artista e Alimentação
  - Avaliação humana (não especialista nos domínios)

Estrutura da apresentação

- Conceitos
- Estado da arte
- Objetivos
  - Contribuições
- Trabalho realizado e em andamento
- Sistema de Extração e Integração de conhecimento geográfico – SEI-Geo
- Avaliação
- Plano, Cronograma e Marcos

Objetivos

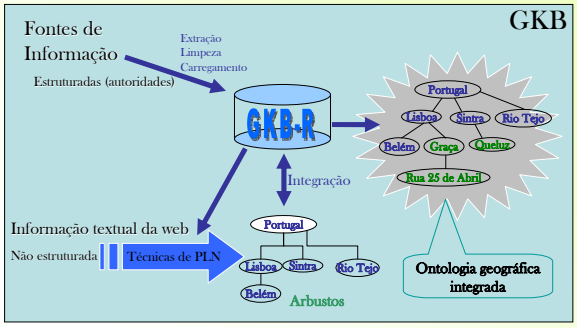
- Dimensionar a "geograficidade" presente em textos da web em português
- Reconhecer o conhecimento disponível em textos e gerar uma floresta de conceitos e ocorrências a partir de textos *(Extração de Informação e Representação de Conhecimento)*
- Integrar a ontologia gerada na Geo-Net-PT

(Integração de Informação e Povoamento e Extensão de Ontologias)

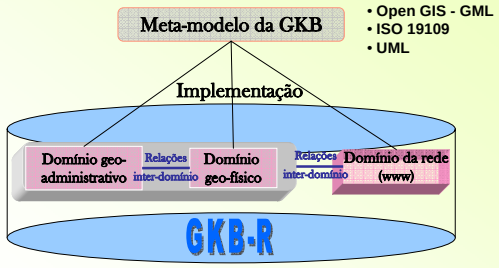
Contribuições

- arquitetura para um sistema de gerenciamento de conhecimento geográfico
- sistema para extração de conhecimento geográfico de textos e integração desse conhecimento em ontologias geográficas, utilizando textos da web em português
- construção e disponibilização pública e gratuita de uma ontologia geográfica de Portugal com conhecimento integrado de diversas fontes e domínios de informação

Ambiente para Gerenciamento de Conhecimento Geográfico

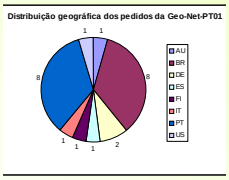


Domínios na GKB



Trabalho realizado e em andamento

- Modelagem e representação de conhecimento geográfico
- Base de dados -> Ontologias
- GKB 1.0 -> Geo-Net-PT01



- GKB-ML
- Geotumba
- Participações no Geo-CLEF 2005 e 2006, HAREM e Mini-HAREM
- GKB 2.0

Principais requisitos do meta-modelo da GKB 2.0

- Suporte para relacionamentos entre tipos
- Suporte para conjunto de atributos genéricos
- Melhor gestão da proveniência da informação
- Suporte multi-língue

Informação geográfica em textos

- Medições Iniciais
  - Conteúdo geográfico dos textos na web portuguesa?
  - Objetivos
    - detecção do conteúdo das EMs geográficas em texto
    - grau de ambigüidade intracategorial (dentro do Geo-Net-PT e dentro das classificações de EM do SIEMÊS e do HAREM
    - sobreposição entre a Geo-Net-PT e os textos da web

Informação geográfica em textos

- Medições Iniciais
  - Amostra de 32.000 documentos etiquetados por um reconhecedor de EMs (SIEMÊS)
  - Categorias: pessoa, organização e local
    - 31% das EMs distintas reconhecidas como pessoa
    - 23% das EMs distintas reconhecidas como organização continham um nome geográfico incluído na Geo-Net-PT
  - Os locais constituem 30% do total de EMs identificadas
  - Considerando apenas os locais distintos, 75% são EMs multi-palavra

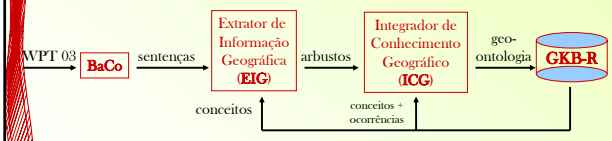
Estrutura da apresentação

- Conceitos
- Estado da arte
- Objetivos
  - Contribuições
- Trabalho realizado e em andamento
- Sistema de Extração e Integração de conhecimento geográfico – SEI-Geo
- Avaliação
- Plano, Cronograma e Marcos

SEI-Geo

- Sistema de Extração e Integração de Conhecimento Geográfico
  - Objetivos
    - reconhecer o conhecimento geográfico disponível em textos
    - gerar uma representação estruturada desse conhecimento
    - integrá-lo no GKB-R

Arquitetura do SEI-Geo



Extrator de Informação Geográfica (EIG)

Algoritmo Inicial: Extração de arbustos

```
1: C = {conceitos da Geo-Net-PT}
2: N = |C|
3: S = {sentenças do BaCo}
4: for all s ∈ S do
5:    C_s = {c ∈ C | ∃c' ∈ s : c' = c}
6:    if |C_s| ≥ 2 then
7:       EM_s = {en ∈ s}
8:       Arbusto = < c, em > |c ∈ C_s, em ∈ s, pos(em, s) = succ(pos(c, s))
9:       Ar_Set = Ar_Set ∪ Arbusto
10:   end if
11: end for
12: for all a ∈ Ar_Set do
13:    conf = confiança(a)
14: end for
```

O Pedro nasceu no **concelho** de Mora, **distrito** de Évora.

distrito de Évora, concelho de Mora

Avaliação

- Extrator de Informação Geográfica
  - Criação de uma "coleção de teste"
  - Precisão e Abrangência
- Integrador de Conhecimento Geográfico
  - Estudos de "mutilação" (*ablation studies*) a partir da Geo-Net-PT

Estado da Arte de Sistemas de EI e II

	Pad	Onto	EEM	ICA	**Web**	**Geo**	PT
Snowball	✓	X	✓	X	**X**	✓	X
KnowItAll	✓	X	✓	X	✓	✓	X
[AM 02]	X	✓ (WordNet)	✓	✓	**X**	**X**	X
[Uryupina 03]	✓	✓ (almanaque)	✓	X	✓	✓	X
OntoLearn	X	✓ (WordNet)	✓	✓	**X**	**X**	X
OntoSyphon	✓	✓	✓	**X**	✓	**X**	**X**
[Borges 06]	✓	X	✓	X	✓	✓	✓
SEI-Geo	✓	✓	✓	✓	✓	✓	✓

- Legenda:
  - PAD = Padrões
  - Onto = Ontologias
  - EEM = Extração de EMs
  - ICA = Integra o Conhecimento Adquirido
  - Web = Textos da Web
  - Geo = Extgrai inforanção geográfica
  - PT = Processa textos em português

Tarefas e Cronograma

1 – Escrita e defesa da proposta
2 – Criação da GKB 2.0
3 – Implementação do Extrator de Informação Geográfica (EIG) (domínio administrativo)
4 – Implementação do Integrador de Conhecimento Geográfico (ICG) (domínio administrativo)
5 – Comparação dos resultados extraídos pelo EIG e pelo ICG com os presentes na GKB 1.0
6 – Participação no Geo-CLEF 2007
7 – Realização das tarefas 3-5 para o domínio físico, sendo que na tarefa 5 será utilizada a GKB 2.0.
8 – Escrita de artigos
9 – Escrita da tese

	2007												2008							
Nº tarefa	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set
1	x																				
2	x																				
3	x	x	x	x	x																
4					x	x	x	x													
5								x	x	x											
6			x	x	x																
7											x	x	x	x							
8				x	x					x	x			x	x						
9																x	x	x	x	x	x

Marcos do plano

- Junho de 2007: lançamento da Geo-Net-PT02 expandida com informação da geografia administrativa e física
- Abril de 2008: lançamento da Geo-Net-PT03 expandida com informação proveniente de textos
- Maio de 2008: versão beta de um protótipo do sistema SEI-Geo

Obrigado pela vossa atenção!!!!!!!!!