
FACULDADE DE CIÊNCIAS UNIVERSIDADE DE LISBOA
www.linguateca.pt

Expansão de ontologia geográfica com textos em Português

Marcirio Silveira Chaves

Orientadores: Diana Santos
Mário J. Silva

Seminário doutoral – Fev./2006

Estrutura da apresentação

- ✓ Motivação
- ✓ Fases do trabalho
- ✓ Trabalho realizado
 - ✓ GKB / Geo-Net-PT01
- ✓ Trabalho em andamento
 - ✓ Hipótese
 - ✓ Objetivos
 - ✓ Experimentos
- ✓ Tarefas a realizar
- ✓ Resumo

23-Fev-06 Seminário doutoral DI-FCUL 2

Web Semântica

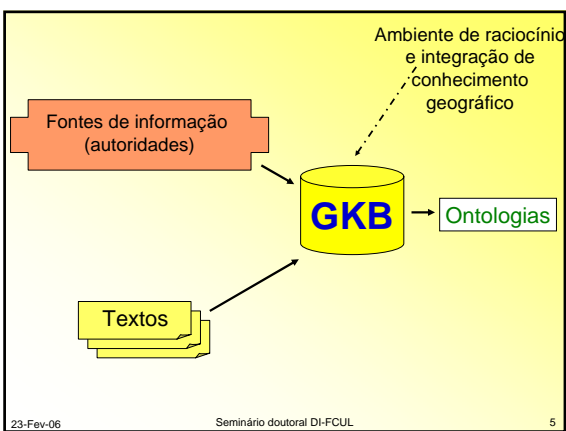
- ✓ 3 visões
 - > IA clássica
 - > Berners-Lee: base de dados
 - > Documentos anotados: PLN como base
- ✓ Ontologia: conceito fundamental na arquitetura da Web Semântica
- ✓ *“Assembling data is no longer the biggest challenge. Instead, the major hurdle these days is one of **data integration**.”*
Russ Altman, Stanford

23-Fev-06 Seminário doutoral DI-FCUL 3

Problema de pesquisa

- ✓ **Coleta, identificação, limpeza, classificação, integração e formalização** da informação geográfica (administrativa) sobre Portugal
 - ✓ Formal
 - ✓ Carência de informação geográfica detalhada formal
 - ✓ Informal
 - ✓ Nomes e relações geográficas informais

23-Fev-06 Seminário doutoral DI-FCUL 4



Fases do trabalho

1. Criação da GKB
2. Caracterização da “geograficidade” existente nos textos
3. Extração de conhecimento geográfico
4. Criação de ontologia geográfica
5. Integração do conhecimento obtido em 4.

23-Fev-06 Seminário doutoral DI-FCUL 6

Contexto

- ✓ Linguateca – Centro de recursos distribuídos para o processamento computacional da língua portuguesa
- ✓ Tumba!
- ✓ Projeto GREASE - *Geographic Reasoning for Search Engines*
- ✓ WPT 03

23-Fev-06

Seminário doutoral DI-FCUL

7

GKB

- ✓ GKB – *Geographic Knowledge Base*
 - KB formada por fontes de informação **distintas, heterogêneas e complementares**
 - Informação geográfica e de rede
 - Mais de 800.000 registos
 - Exportada como ontologias
 - Geo-Net-PT01
- Feature
 - Um objeto com significado no domínio selecionado do discurso [ISO19109].
 - Ex.: países, cidades e localidades

23-Fev-06

Seminário doutoral DI-FCUL

8

Integração de conhecimento na GKB

- ✓ Hierarquia da GKB composta a partir de diferentes fontes de informação
- ✓ Algoritmo:
 - Procura o **tipo de feature comum em mais baixo nível** em ambas as hierarquias
 - Se encontra, ele identifica as **ocorrências comuns entre as hierarquias**
 - Sobe na hierarquia e procura pelo **ascendente comum em mais baixo nível**
 - Verifica a distância (em nº de relacionamentos parteDe) entre as ocorrências comuns dos tipos de *features* e seus ascendentes. O ascendente que tem a menor distância até as ocorrências comuns é **integrado** com um relacionamento parteDe com o ascendente na outra hierarquia

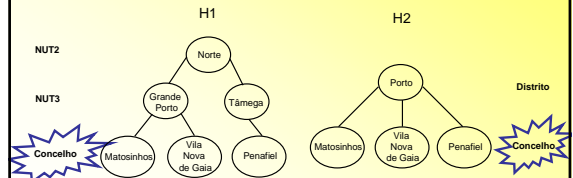
23-Fev-06

Seminário doutoral DI-FCUL

9

Integração de conhecimento na GKB

- ✓ Hierarquia da GKB composta a partir de diferentes fontes de informação



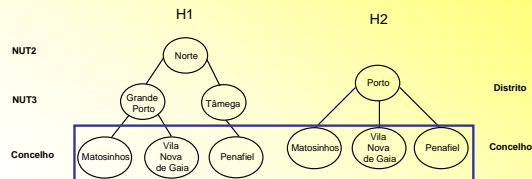
23-Fev-06

Seminário doutoral DI-FCUL

10

Integração de conhecimento na GKB

- ✓ Hierarquia da GKB composta a partir de diferentes fontes de informação



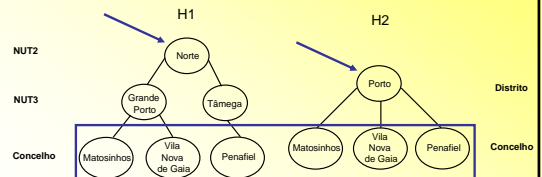
23-Fev-06

Seminário doutoral DI-FCUL

11

Integração de conhecimento na GKB

- ✓ Hierarquia da GKB composta a partir de diferentes fontes de informação

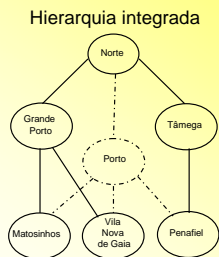


23-Fev-06

Seminário doutoral DI-FCUL

12

Integração de conhecimento na GKB



➤ Os relacionamentos existentes em ambas as hierarquias são mantidos.

23-Fev-06

Seminário doutoral DI-FCUL

13

Inserindo âmbitos geográficos na GKB

- ✓ Âmbitos geográficos
 - www.cm-lisboa.pt
 - Lisboa (concelho)
- ✓ Fatos e regras
- ✓ Novos relacionamentos e conhecimento
- ✓ Lógicas de Descrição (LDs)
- ✓ Domínio geográfico
 - Nomes compostos por múltiplas palavras são representados de diferentes formas
- ✓ Domínio de rede
 - Nomes de URLs são decompostos com base na divisão de domínios correspondente

23-Fev-06

Seminário doutoral DI-FCUL

14

Inserindo âmbitos geográficos na GKB

- ✓ ABox em LDs para o:
 - Concelho de **Santiago do Cacém**
 - `geoFeatureName(270,"santiagodocacem").`
 - `geoFeatureName(270,"santiogocacem").`
 - `geoFeatureName(270,"santiago-do-cacem").`
 - `geoFeatureName(270,"santiago-cacem").`
 - `geoFeatureType(270,"CON").`
 - Web site: **www.cm-santiago-do-cacem.pt**
 - `netSiteSubDomain(33684,"www").`
 - `netSitePrefix(33684,"cm").`
 - `netSiteDomainToken(33684,"santiago-do-cacem").`
 - `netSiteTLD(33684,"pt").`

23-Fev-06

Seminário doutoral DI-FCUL

15

Inserindo âmbitos geográficos na GKB

- ✓ Descrição da terminologia (TBox em LDs)
 - **Concelhos**

$$\text{hasScope}(\text{idN}, \text{idG}) \equiv$$

$$\exists \text{netSiteDomainToken}(\text{idN}, X) \cap$$

$$((\exists \text{netSitePrefix}(\text{idN}, "cm") \cup \exists \text{netSitePrefix}(\text{idN}, "mun"))) \cap$$

$$\exists \text{geoFeatureType}(\text{idG}, "CON") \cap$$

$$\exists \text{geoFeatureName}(\text{idG}, X).$$

23-Fev-06

Seminário doutoral DI-FCUL

16

Inserindo âmbitos geográficos na GKB

- ✓ Ex.:

$$\text{hasScope}(\text{idN}, \text{idG}) \equiv$$

$$\exists \text{netSiteDomainToken}(\text{idN}, X) \cap$$

$$((\exists \text{netSitePrefix}(\text{idN}, "cm") \cup \exists \text{netSitePrefix}(\text{idN}, "mun"))) \cap$$

$$\exists \text{geoFeatureType}(\text{idG}, "CON") \cap$$

$$\exists \text{geoFeatureName}(\text{idG}, X).$$
 - `netSiteDomainToken(33684, "santiago-do-cacem").`
 - `netSitePrefix(33684, "cm").`
 - `geoFeatureType(270, "CON").`
 - `geoFeatureName(270, "santiago-do-cacem").`

Novo conhecimento: `hasScope(33684, 270)`.

23-Fev-06

Seminário doutoral DI-FCUL

17

Inserindo âmbitos geográficos na GKB

- ✓ Âmbitos atribuídos aos sites de Portugal com base nas regras

Tipo de Site	# de domínios	# de combinações de sites
distritos	33	17 (52%)
concelhos	288	261 (90%)
freguesias	300	124 (41%)
escolas básicas	1955	124 (6%)
centros de formação	152	55 (36%)
escolas secundárias	402	105 (26%)

- ✓ Âmbitos são estendidos para as páginas web abaixo de cada site com âmbito atribuído

23-Fev-06

Seminário doutoral DI-FCUL

18

Trabalho em andamento

Hipótese:

Existe informação geográfica relevante e interessante na web e é possível integrá-la em ontologias geográficas

Objetivos subsequentes

- ✓ explorar a informação geográfica em textos em português de forma a suportar a expansão de conceitos, relações e ocorrências de uma ontologia geográfica
 - caracterizar a informação geográfica presente em textos na web portuguesa
 - extrair fatos e relações geográficas
 - comparar uma ontologia geográfica derivada de textos em linguagem natural com uma criada a partir de fontes de dados administrativas
 - integrar fatos e relações geográficas na ontologia

Caracterização da informação geográfica presente em textos na web portuguesa

✓ Objetivo

- Ter uma idéia preliminar da informação geográfica na web portuguesa
- Verificar a frequência e distribuição das EMs (SER, ORG, LOC)
- Comparação com o conhecimento armazenado da GKB

1º Estudo

- ✓ SIEMÊS – sistema REM
- ✓ 1.000 documentos (1.704.679 palavras)
- ✓ Cada documento apresenta, em média:
 - 15,6 EMs distintas
 - 3,8 EMs geográficas distintas
- ✓ 60,58% das EM distintas são multi-palavra.
- ✓ 50% das EMs geográficas distintas são multi-palavra.

1º Estudo

- ✓ Nomes de cidade, município ou vila (POV)
 - tipo de geo-EM mais frequente (77%)
- ✓ Média: quase 3 EMs geográficas POV por documento
- ✓ 80% das geo-EMs na amostra analisada **não** estão incluídas na Geo-Net-PT01
- ✓ 21.92% das localidades identificadas pelo SIEMÊS são nomes de localidades considerados oficiais pelas fontes de informação administrativas portuguesas
- ✓ A parte da web portuguesa relacionada com Portugal é pequena (~20%).

2º Estudo

- ✓ Relação entre as categorias SER, ORG e LOC
- ✓ 5.500 documentos aleatórios
- ✓ Em média ...
 - cada documento contém 1.562 palavras
 - 1 EM da categoria Pessoa em cada 245 palavras
 - 1 Organização em cada 248 palavras
 - 1 Localidade em cada 358 palavras

2º Estudo

- ✓ Amostra: 5 grupos de 1.000 documentos
- ✓ EMs da categoria Local do tipo POV em comparação com a Geo-Net-PT01, somente de 19% a 22% das localidades extraídas de texto estão na Geo-Net-PT01
- ✓ 5% dos nomes de pessoas e organizações são idênticos aos nomes de localidades
- ✓ 46% dos nomes multi-palavra de pessoas e 51% dos nomes multi-palavra de organizações contêm um nome geográfico

23-Fev-06

Seminário doutoral DI-FCUL

31

3º Estudo

- ✓ 30.000 documentos
- ✓ 190 Mbytes de texto anotado
- ✓ 201.691 EMs distintas correspondentes a três categorias: SER, ORG, LOC
- ✓ Média
 - ✓ 6,72 EMs distintas por documento
 - ✓ 1,33 LOCs por documento

23-Fev-06

Seminário doutoral DI-FCUL

32

3º Estudo

- ✓ Das 40.022 localidades distintas, existem 27.463 localidades do tipo POV, das quais **5.140 (18,7%) estão na Geo-Net-PT01**
 - Geografia física
 - Nomes de fora de Portugal
 - Nomes informais
- ✓ + de 68% da localidades geográficas detectadas pelo SIEMÊS são do tipo POV
- ✓ Ambigüidade
 - 63.2% dos nomes de pessoas e
 - 54.5% das organizações contêm um nome geográfico

23-Fev-06

Seminário doutoral DI-FCUL

33

Análise dos experimentos

# docs.	36.500
Média EM distintas p/ doc.	6,48
Média LOC distintas p/ doc.	1,30
EM distintas multi-palavra (%)	77,80
Localidades (POV) na Geo-Net-PT01 (%)	15,8

- ✓ LOCs representam cerca de 20% das EMs na amostra analisada do WPT 03
- ✓ EM distintas multi-palavra são pervasivas
- ✓ ~85% das localidades do tipo POV não estão na Geo-Net-PT01

23-Fev-06

Seminário doutoral DI-FCUL

34

Tarefas a realizar

- ✓ Extração de informação em textos web
 - Relações semânticas existentes na Geo-Net-PT-01
 - Parte de, adjacente, equivalente
 - Relações entre categorias
 - LOC-LOC
 - ORG-LOC
- ✓ Comparação de uma ontologia geográfica derivada de textos em linguagem natural com uma criada a partir de fontes de dados administrativas
- ✓ Integração de fatos e relações geográficas na ontologia

23-Fev-06

Seminário doutoral DI-FCUL

35

Resumo

- ✓ GKB – Geo-Net-PT01
- ✓ Análise da geograficidade da web portuguesa
 - Experimentos com sistemas REM
- ✓ Extração de Informação geográfica
- ✓ Expansão da Geo-Net-PT01 com conteúdo de textos da web portuguesa

23-Fev-06

Seminário doutoral DI-FCUL

36

Resultados Parciais

✓ Teóricos

- Chaves, Marcirio Silveira; Santos, Diana. *What kinds of geographical information are there in the Portuguese Web?* PROPOR, 2006. (no prelo)
- Chaves, Marcirio Silveira; Silva, Mário J. e Martins, Bruno. *A Geographic Knowledge Base for Semantic Web Applications*. SBB05, pp. 40-54, 2005.
- Chaves, Marcirio Silveira; Silva, Mário J. e Martins, Bruno. *GKB - Geographic Knowledge Base*. DI/FCUL, TR05-12, Julho, 2005.
- Martins, Bruno, Chaves, Marcirio Silveira e Silva, Mário J. *Assigning Geographical Scopes To Web Pages*. ECIR 2005: 564-567, 2005
- Martins, Bruno, Chaves, Marcirio Silveira e Silva, Mário J. *Challenges and resources for evaluating geographical IR*. GIR 2005: 65-69, 2005.

Resultados Parciais

✓ Teóricos

- Silva, Mário J.; Martins, Bruno; Chaves, Marcirio Silveira; Cardoso, Nuno; Afonso, Ana Paula. *Adding Geographic Scopes to Web Resources*. CEUS - Computers, Environment and Urban Systems, Elsevier Science. (no prelo).
- Cardoso, Nuno; Martins, Bruno; Chaves, Marcirio Silveira; Andrade, Leonardo; Silva, Mário J. *The XLDB Group at GeoCLEF 2005*. 6th CLEF Workshop, 2005.
- Santos, Diana et al. *Linguatca: um Centro de Recursos Distribuído para o Processamento Computacional da Língua Portuguesa*. Proc. of the international workshop "Taller de Herramientas y Recursos Lingüísticos para el Español y el Portugués", pp. 147-154, IBERAMIA, Puebla, Mexico, 2004.

✓ Práticos

- **Geo-Net-PT01**: Primeira ontologia geográfica pública de Portugal - <http://xldb.di.fc.ul.pt/geonetpt>