

UNIVERSIDADE DE LISBOA

Faculdade de Ciências

Departamento de Informática



MEDIDAS DE SEMELHANÇA SEMÂNTICA
APLICADAS ÀS
ONTOLOGIAS GEOGRÁFICAS

Daniel António Correia Amoedo

MESTRADO EM ENGENHARIA INFORMÁTICA

Especialização em Sistemas de Informação

2011

UNIVERSIDADE DE LISBOA

Faculdade de Ciências

Departamento de Informática



MEDIDAS DE SEMELHANÇA SEMÂNTICA
APLICADAS ÀS
ONTOLOGIAS GEOGRÁFICAS

Daniel António Correia Amoedo

Projecto

Trabalho orientado pelo Prof. Doutor Francisco José Moreira Couto

MESTRADO EM ENGENHARIA INFORMÁTICA

Especialização em Sistemas de Informação

2011

Agradecimentos

Ao Prof. Doutor Francisco José Moreira Couto, pela sua orientação, conhecimentos, entusiasmo, disponibilidade e apoio durante a realização do trabalho, essenciais para o seu sucesso.

À excepcional equipa que constitui o grupo de pesquisa do LASIGE, pela sua organização, pela informação disponibilizada, pelas reuniões constantes das quais constituíram também elas uma ajuda imprescindível à realização deste projecto.

Os meus agradecimentos à FCT (Fundação para a Ciência e Tecnologia) através do projecto PTDC/EIA/73614/2006, cc 4895 por ter financiado o trabalho de investigação apresentado neste relatório.

À minha família pela sua ajuda e apoio incondicional durante toda a minha fase académica, e por sempre acreditarem em mim e nas minhas capacidades.

A todos os meus amigos pela sua ajuda, ideias e pelas palavras acertadas que me ajudaram a superar as adversidades exteriores ao projecto.

Resumo

É cada vez mais recorrente o uso da Internet na procura de informação específica, sendo que muitas vezes essa procura assenta em contextos geográficos. A presente tese descreve o trabalho desenvolvido no âmbito do projecto GREASE, que estuda métodos de recuperação e extracção de informação geográfica para grandes colecções de texto, com ênfase na Web.

Este trabalho baseou-se na implementação de várias medidas de semelhança semântica aplicadas às ontologias geográficas. Estas medidas foram anteriormente desenvolvidas e aplicadas no âmbito da Linguagem Natural e da Bioinformática. O objectivo desta tese passou igualmente pela aplicação destas medidas no desenvolvimento de estratégias de desambiguação de termos geográficos que partilham o mesmo nome.

Elaborou-se, ainda, um estudo que pretendeu averiguar, entre as medidas de semelhança semântica, quais as que melhor se adaptariam a uma ontologia geográfica. Sugeriu-se que as medidas mais eficazes são aquelas que usam o MICA ou o GRASM para fazer diferenciação entre pares de termos de valor de conteúdo de informação semelhantes. Adicionalmente, foi realizada a desambiguação de referências geográficas extraídas de um “site” da Web através de conhecimento ontológico e do uso das medidas de semelhança semântica.

Palavras-chave: Medidas de Semelhança Semântica, Ontologias Geográficas, Desambiguação de Termos, Contextos Geográficos

Abstract

The use of the Internet in the search of specific information is increasingly recurrent and often based on geographical contexts. The information on the Web is vast and scattered and giving a meaning to that information through the use of ontologies has been a natural evolution of the research methods.

This thesis describes the work developed under the GREASE project, which studies methods of Web based information retrieval and extraction for large collections of text with emphasis on the Web.

The work that has been carried out analyses the implementation of several semantic similarity measures that were previously developed under the Natural Language and the Bioinformatics research in order to be applied to geographic ontologies. Therefore, we developed strategies for disambiguation of geographical terms that share the same name. Moreover, a study on how to find which of the semantic similarity methods can better adapt to a geographical ontology was also carried out.

The results of this study suggest that measures of semantic similarity that better fit the geographical ontology used in the project GREASE are those that use MICA or GRASM approach to differentiate between pairs of terms with similar value of information content. Additionally, it was also possible to perform disambiguation of geographical references from a Web site using ontological knowledge and the use of Semantic Similarity Measures.

Keywords: Semantic Similarity Semantic Similarity Measures, Geographic Ontology, Disambiguation Terms and Geographical Contexts

Conteúdo

Capítulo 1	Introdução.....	11
1.1	Motivação.....	11
1.2	Objectivos.....	13
1.3	Metodologia.....	14
1.4	Estrutura do documento.....	15
Capítulo 2	Conceitos e Trabalho Relacionado.....	17
2.1	Terminologia.....	17
2.2	Ontologias.....	18
2.3	Similaridade entre grafos.....	19
2.4	Similaridade Semântica.....	20
2.4.1	Medidas baseadas no comprimento do caminho.....	20
2.4.2	Medidas baseadas no Conteúdo de informação.....	21
2.4.3	GRASM - Graph-Based Similarity Measure.....	23
2.4.4	Medida Híbrida.....	23
2.4.5	GKB - Geographic Knowledge Base.....	24
2.4.6	Geo-Net-PT 01.....	26
2.4.7	Geo-Net-PT 02.....	29
2.5	Web 1T 5-gram Corpus (Google N-Grams).....	30
2.6	jWeb1T.....	31
Capítulo 3	GeoSSM.....	33
3.1	Arquitectura.....	33
3.2	Alterações à Geo-Net-PT.....	34
3.2.1	Tabela SSM_GraphPath.....	35

3.2.2	Tabela SSM_TermFreq	37
3.2.3	Estrutura	39
3.3	Exemplos	41
3.3.1	1º Exemplo de SSM (MICA igual)	41
3.3.2	2º Exemplo de SSM (MICA e IC igual)	43
3.3.3	3º Exemplo de SSM (MICA e antecessores comuns iguais).....	44
3.4	Medida Combinada - SimGeo	45
3.4.1	Exemplo de Medidas de Semelhança Semântica com a SimGeo	46
Capítulo 4	GeoScope - Geographical Scope.....	49
4.1	Desambiguação de Referências Geográficas.....	50
4.1.1	Desambiguar uma referência geográfica.....	51
4.1.2	Desambiguação de um conjunto de referências geográficas.....	53
4.1.3	Desambiguação de nomes da Geo-Net-PT: exemplo.....	55
4.2	Exemplos	63
4.2.1	Utilização do trabalho desenvolvido.....	65
Capítulo 5	Conclusões	69
5.1	Experiências com as várias Medidas de Semelhança Semântica	70
5.2	Desambiguação do Âmbito Geográfico	70
5.3	Geo-Net-PT	71
5.4	Trabalho Futuro	72
Bibliografia	75

Lista de Figuras

Figura 2.1: Esquema para o cálculo do IC	21
Figura 2.2: Meta-modelo do GKB	25
Figura 2.3: Relações entre tipos de conceitos para os dados da Geo-Net-PT	27
Figura 2.4: Exemplo de uma feature na Geo-Net-PT 01	28
Figura 3.1: Arquitectura da GeoSSM.	33
Figura 3.2: Rrelação de “part-of” entre o termo 146 (Lisboa) e os seus antecessores “adm_feature_relationship”	36
Figura 3.3: Relação de “part-of” entre o termo 146 (Lisboa) e os seus antecessores que pode ser observada na tabela "ssm_graphpath".	36
Figura 3.4: Relações do tipo “Part of” para os dados físicos da Geo-Net-PT.	40
Figura 3.5: Subgrafo exemplificativo de “paths” entre termos.....	42
Figura 4.1: Valores de IC para termos de nome ‘Lisboa’ na Geo-Net-PT.	51

Lista de Tabelas

Tabela 2.1:Distribuição e ambiguidade dos termos na Geo-Net-PT	27
Tabela 2.2: Diferenças entre a Geo-Net-PT 2.1 e a Geo-Net-PT 2.2.	29
Tabela 2.3:Quantificação dos nomes da Geo-Net-PT 02 no Google N-Grams	30
Tabela 3.1:Esquema da Tabela "ssm_graphpath"	35
Tabela 3.2: Antecessores do termo ID:146 na tabela “adm_feature_relationship”	35
Tabela 3.3: Antecessores do termo ID:146 na tabela “ssm_graphpath”	36
Tabela 3.4 Tabela " ssm_termfreq " da base de dados Geo-Net-PT-SSM.	37
Tabela 3.5: Percentagem de features com o mesmo valor de IC na Geo-Net-PT...	41
Tabela 3.6: Termos utilizados no 1º exemplo de SSM	41
Tabela 3.7: Semelhança semântica entre os termos 182631, 197748 e 224027	42
Tabela 3.8: Termos utilizados no 2º exemplo de SSM	43
Tabela 3.9: Semelhança semântica entre os termos 418454, 3646 e 2178.	44
Tabela 3.10: Termos utilizados no 3º exemplo de SSM	44
Tabela 3.11: Semelhança semântica entre os termos 1704, 1705 e 1709.	45
Tabela 3.12: Termos utilizados no 4º exemplo de SSM.	46
Tabela 3.13: Semelhança semântica na SimGeo.	46
Tabela 4.1: Termos utilizados nas exemplificações do Capítulo 4 (GeoScope)	55
Tabela 4.2: EDNG1 - Combinações de termos utilizados neste exemplo	56
Tabela 4.3:EDNG2 - Semelhança semântica entre termos das duas primeiras referências geográficas fixadas pela heurística.	57
Tabela 4.4: EDNG2 - Semelhança semântica entre os termos referentes a “Praça Afonso de Albuquerque” e o âmbito encontrado na iteração anterior.	57
Tabela 4.5: EDNG3 - Âmbito geográfico ontológico escolhido pela heurística. ...	58
Tabela 4.6: EDNG3 - Semelhança semântica entre os termos e âmbito geográfico calculado pela heurística.	58
Tabela 4.7: EDNG4 - Conjunto de termos candidatos a âmbito geográfico.	59

Tabela 4.8: EDNG4 - Semelhança semântica entre os termos das referências geográficas e o âmbito ontológico 129.....	59
Tabela 4.9: EDNG4 - Semelhança semântica entre os termos das referências geográficas e o âmbito ontológico 3965.....	60
Tabela 4.10: EDNG4 – Média de semelhança semântica entre termos de cada conjunto criado.	60
Tabela 4.11: EDNG5 - Conjunto de termos candidatos a âmbito geográfico.	61
Tabela 4.12: EDNG5 – Semelhança semântica entre os pares de termos gerados por EDNG1 para os sucessores do termo 129.	61
Tabela 4.13: EDNG5 – Semelhança semântica entre os pares de termos gerados por EDNG1 para os sucessores do termo 129.	62
Tabela 4.14: EDNG5 – Médias de semelhança semântica entre os termos dos conjuntos produzidos.....	62
Tabela 4.15: Referências geográficas extraídas manualmente das páginas Web do site Portugal Tribe.	63
Tabela 4.16: Referências ontológicas e respectivos âmbitos geográficos ontológicos calculados pelas SSM com as diferentes EDNG	64
Tabela 4.17: Âmbitos Geográficos Ontológicos das páginas do site Portugal Tribe e respectivos âmbitos geográficos ontológicos calculados.	65
Tabela 4.18: Percentagem de entidades correctamente extraídas e percentagem das entidades correctamente desambiguadas.	66

Capítulo 1 Introdução

As pesquisas na Web desconhecem, actualmente, o significado das palavras contidas nos documentos. Os modelos clássicos de recuperação de informação consideram que cada documento é representado por um conjunto de palavras, devolvendo-se o resultado escolhido de acordo com um peso atribuído a uma palavra num documento e a um “ranking” associado a esse documento. A utilização de ontologias desempenha, então, um papel importante para que o sistema computacional possa entender esse significado; permite, por exemplo, efectuar uma representação de um conjunto de conceitos dentro de um domínio, dando, a um sistema computacional, uma compreensão mais alargada do mundo (ou, pelo menos, do mundo que é representado na ontologia através do relacionamento entre conceitos.). A comparação entre conceitos constitui uma das principais operações de que a utilização de ontologias por um sistema computacional pode vir a revelar-se vantajosa. Esta comparação torna-se possível graças à implementação de medidas de semelhança semântica, avaliando-se o grau de similaridade entre dois conceitos organizados numa ontologia. Medidas, aliás, que no passado foram desenvolvidas e aplicadas na Linguagem Natural e na Bioinformática.

1.1 Motivação

O principal pressuposto do GREASE-I, bem como o de vários trabalhos de investigação que o antecederam, pode ser descrito como o facto de o âmbito geográfico de um documento poder ser representado por um único conceito ou forma geográfica. Esta última pode, por sua vez, ser obtida a partir dos meta-dados ou através de técnicas de extracção de informação aplicadas ao texto e de algoritmos de grafos aplicados aos seus “links”.

O GREASE-II¹ rompe, contudo, com este pressuposto. O âmbito geográfico de um documento é neste caso caracterizado por um resumo constituído por um conjunto de etiquetas geográficas. Tal como nas aplicações da Web2.0, algumas das etiquetas irão corresponder directamente a conceitos ontológicos geo-referenciados, enquanto outras etiquetas podem não ser geo-referenciáveis. No GREASE-II (<http://xldb.fc.ul.pt/wiki/Grease>) presume-se que estes resumos (que também podem ser gerados a partir dos conteúdos por métodos de extracção de informação, de aprendizagem ou por algoritmos de grafos), constituem uma descrição muito mais rica do teor dos conteúdos em termos geográficos. Através do uso destes resumos, as pesquisas geográficas melhoram significativamente, permitindo representar a área de interesse geográfico dos documentos, contornando-se muitas das limitações impostas pelo modelo simples, e baseado em âmbitos geográficos, que foi aprofundado no GREASE-I.

Uma pesquisa geográfica que seja efectuada num Recuperador de Informação Geográfica (GIR) não é mais do que uma pesquisa num Recuperador de Informação (IR) dito normal, ao qual foram adicionados dados geográficos; esse acréscimo permite, aos utilizadores, a pesquisa de documentos correspondentes a âmbitos geográficos de um dado lugar, originando, até, uma maior especialização na pesquisa de informação em documentos relevantes para esse domínio (Martins, 2008). Exige-se, aos novos sistemas, que conheçam especificamente o contexto geográfico dos documentos, atendendo a que o seu propósito é devolver documentos geo-referenciados que possam ser relevantes para a região referida na consulta efectuada. “Os resultados gerados pelos IR podem ser classificados com alguma medida que meça a relevância geográfica de um documento” (Martins, 2008).

Podemos afirmar que uma ontologia geográfica é uma representação do conhecimento de como o nosso mundo, ou parte dele, expressa as relações existentes entre as várias zonas geográficas, relacionando-se, estas, com os diferentes modos de divisão do mundo e permitindo, a um sistema computacional, a inferência sobre aquele. Para que o sistema computacional perceba o contexto geográfico de um documento, torna-se necessário que as suas etiquetas refiram termos da ontologia; cria-se, então, a necessidade de desenvolver métodos que efectuem esse reconhecimento e mapeamento de uma forma automática. Dado que o contexto geográfico é incorporado através de descrições de linguagem natural, geram-se problemas ao nível dos nomes, que podem facilmente ser ambíguos (Martins, 2009).

¹ <http://xldb.fc.ul.pt/wiki/Grease>

É essencial que se proceda à desambiguação desses nomes, tornando possível a percepção pelo computador do verdadeiro âmbito geográfico que representam as referências geográficas contidas nos documentos. Só assim, reconhecido esse âmbito geográfico, é que o computador devolverá resultados que, com maior probabilidade, se enquadrarão com o âmbito da pesquisa. A apreensão do âmbito geográfico de um documento permitirá inferir quais os termos da ontologia que expressam os locais das referências geográficas contidas nesse documento através da similaridade com aquele. Uma forma de calcular essa similaridade é utilizar medidas que calculem a similaridade entre os termos da ontologia, tal como as medidas de similaridade entre grafos e as medidas de semelhança semântica (Resnik, 1995).

O presente trabalho está inserido no projecto GREASE-II, o qual tem como objectivo dar continuidade à investigação inicial desenvolvida no GREASE-I, focada em algoritmos de atribuição de âmbitos geográficos a conteúdos da Web e o desenvolvimento de novos métodos de recuperação de informação geográfica baseada nos âmbitos atribuídos.

Foram implementadas medidas de semelhança semântica no contexto do GIR que calculam a semelhança semântica entre termos num âmbito geográfico, e desenvolveram-se métodos de desambiguação de referências geográficas através dessas medidas.

1.2 Objectivos

O principal objectivo do presente trabalho consiste na implementação e no estudo de medidas de semelhança semântica no âmbito da Geo-Net-PT, uma ontologia geográfica de Portugal (Lopez et al., 2009). Estas medidas foram utilizadas no cálculo da semelhança entre conjuntos de termos geográficos. O processo dividiu-se em três partes:

- Cálculo da similaridade entre dois termos da Geo-Net-PT: as medidas de semelhança semântica calculam a similaridade entre conceitos que estejam organizados numa ontologia. Aplicando este tipo de medidas a uma ontologia geográfica, obter-se-á uma produção de valores quantitativos da similaridade entre dois locais associados ao âmbito geográfico descrito pela ontologia.

- Desambiguação das Referências Geográficas: nos casos em que se verifique a existência, em Portugal, de vários locais com o mesmo nome, a ontologia que o descreve também conterá termos representados por nomes iguais. Daqui resultará a necessidade de efectuar a desambiguação das referências geográficas existentes nos resumos geográficos, obtendo-se, assim, o correcto mapeamento daquelas de acordo com os termos da ontologia. O processo de desambiguação não trata, de forma isolada, uma referência geográfica. Antes irá mapear, coerentemente, as referências geográficas no contexto de um conjunto de termos da ontologia; por outras palavras, assume que as referências geográficas, num mesmo contexto, devem ser similares entre si. A desambiguação irá produzir resumos de entidades geográficas reconhecidas numa ontologia, representando-as através dos seus identificadores nessa ontologia.

Pretende-se, através do estudo destes conceitos, melhorar a resposta de pesquisas em sistemas de recuperação de informação que contenham referências geográficas, procurando obter uma utilização mais eficaz das referências geográficas contidas em documentos.

1.3 Metodologia

Dividiu-se, em quatro fases distintas, a implementação das medidas de semelhança semântica no contexto geográfico:

- Análise das medidas de semelhança semântica criadas no âmbito da WordNet e da Bioinformática;
- Análise dos dados na Geo-Net-PT, uma ontologia geográfica pública e de âmbito nacional, que contém dados administrativos sobre distritos, concelhos, freguesias, assim como as suas relações, dados populacionais e coordenadas geográficas;
- Desenvolvimento do software GeoSSM que implementa as medidas de semelhança semântica para a ontologia geográfica Geo-Net-PT;

- Desenvolvimento do software GeoScope que desambigua referências geográficas ambíguas na Geo-Net-PT e atribui o âmbito geográfico a esse conjunto de referências geográficas;

1.4 Estrutura do documento

A estrutura do documento obedece a uma divisão da análise por cinco capítulos. O presente capítulo efectua uma breve introdução ao projecto, delineando a exposição de motivos e os objectivos que o justificaram. De seguida, encontramos, nos conceitos e trabalho relacionado, uma introdução aos conceitos que serviram de base à apresentação da tese e uma descrição das tecnologias utilizadas.

No capítulo terceiro, sob o título GeoSSM, formula-se uma metodologia de aplicação das medidas semânticas a ontologias geográficas, nomeadamente à Geo-Net-PT, uma ontologia descritiva da organização territorial portuguesa. É ainda apresentado o conjunto de testes efectuados às várias medidas de semelhança semântica implementadas e os respectivos resultados obtidos, sugerindo-se uma proposta de produzir semelhança semântica entre os termos da Geo-Net-PT.

O quarto capítulo, de título GeoScope, é dedicado à exposição do sistema desenvolvido para desambiguar termos da ontologia geográfica usada, efectuando-se, esse processo de desambiguação, a partir das medidas de semelhança semântica implementadas; por outro lado, o sistema serve o propósito de definir o âmbito geográfico de um conjunto de referências geográficas. Por fim, são apresentados vários testes que validam a importância da utilização das medidas de semelhança semântica no âmbito das pesquisas geográficas através da desambiguação de termos pelo uso das medidas implementadas.

O último capítulo efectua uma análise crítica do trabalho desenvolvido e dos resultados obtidos, propondo-se, também, algumas sugestões de trabalho a desenvolver no futuro.

Capítulo 2 Conceitos e Trabalho

Relacionado

Neste capítulo descrevem-se os conceitos que serviram de base a esta tese e, sucintamente, far-se-á uma apresentação das principais tecnologias utilizadas.

Em primeiro lugar, importa referir a importância do uso das ontologias e das medidas de semelhança semântica e o benefício que daí pode resultar para a pesquisa geográfica.

Será ainda apresentado um “corpus”, denominado Google N-Grams, capaz de satisfazer as necessidades associadas ao cálculo de semelhança semântica entre termos contidos na ontologia, atendendo à aplicação de medidas de semelhança semântica. Esta ontologia e este “corpus” são fundamentais para a implementação da ferramenta que foi desenvolvida para o cálculo da similaridade entre locais no território português.

2.1 Terminologia

As terminologias utilizadas ao longo desta tese baseiam-se em trabalhos anteriormente desenvolvidos no âmbito do projecto GREASE, incidindo maioritariamente nos trabalhos relacionados com Recuperação de Informação Geográfica (Chaves et al., 2008) e com Medidas de Semelhança Semântica (Pesquita et al. 2009).

Como tal, aplicaram-se as terminologias abaixo identificadas:

- Referência Geográfica (RG): todo o nome próprio que refere um local geográfico. É uma designação equivalente a Nome de Local.

- Referência Ontológica (RO): é um termo geográfico definido, sem qualquer ambiguidade, por um único identificador, equivalendo a um termo no âmbito da Geo-Net-PT.
- Esboço de Entidade (EE): representa todas as entidades que têm um mesmo nome. Por exemplo, na Geo-Net-PT existem 207 ROs que têm nome igual a “Rua Vasco da Gama”.
- Âmbito Geográfico Ontológico (AGO): é o antecessor comum mais informativo na Geo-Net-PT a um dado conjunto de RO.
- Medidas de Semelhança Semântica (SSM) – calculam a similaridade entre termos que estejam organizados numa ontologia.

2.2 Ontologias

O vocábulo “ontologia” foi introduzido no ramo da Inteligência Artificial por Grubber, que o definiu como sendo uma especificação de conceptualizações. As ontologias são utilizadas em sistemas baseados no conhecimento, já que disponibilizam uma estrutura declarativa que é utilizada na tentativa de se obter uma inferência automática. O uso de ontologias permite a partilha de informação entre seres humanos e programas (Gruber, 1991).

Um corpo de conhecimento formalmente representado baseia-se numa conceptualização de objectos, conceitos e outras entidades, bem como no relacionamento entre estes. Trata-se, basicamente, de uma representação abstracta e simplificada da realidade. Assim, qualquer base de conhecimento, qualquer sistema baseado no conhecimento ou mesmo qualquer nível de conhecimento, implica, de forma implícita ou explícita, a construção de uma conceptualização.

Uma ontologia é uma especificação explícita de uma conceptualização e tem por objectivo estabelecer uma boa representação do conhecimento, organizando informação não estruturada.

As ontologias têm inúmeras utilidades, e tanto poderão aplicar-se à organização de sítios da “Internet”, como se revelam essenciais na ajuda à navegação, na ajuda das anotações nas páginas electrónicas (ao disponibilizar uma classificação semântica) e ainda no suporte às pesquisas de informação. Este suporte assume duas formas: a) expande a procura, ao combinar um termo indicado pelo utilizador com termos alternativos na ontologia e b) restringe a busca ao retirar qualquer ambiguidade ao termo, conseguindo, desta forma, uma identificação mais exacta do termo procurado pelo utilizador a partir do termo inicialmente introduzido na pesquisa (Chaves et al., 2009). Desta forma, numa pesquisa geográfica que utilize uma abordagem que recorra ao uso das ontologias, haverá uma maior qualidade nos resultados das pesquisas e estes irão conter informação mais relevante por comparação com uma pesquisa que meramente recorra a meios sintácticos (Fonseca, 2002).

Na desambiguação de termos de uma ontologia geográfica, utilizam-se as referências geográficas que correspondem a âmbitos geográficos mais alargados, o que origina que sejam desambiguados os termos mais específicos. Essa desambiguação pode ser conseguida recorrendo ao uso de medidas que calculem a similaridade entre estes termos, seleccionando aqueles cuja similaridade é mais acentuada.

A ontologia utilizada neste projecto, a Geo-Net-Pt, representa um nível de abstracção de dados geográficos referentes a locais situados em Portugal (Distritos, Freguesias, Arruamentos, etc.), descrevendo a hierarquia e as relações existentes entre esses dados. A modelação dos dados centra-se nas diferentes formas como Portugal está organizado territorialmente e, ao mesmo tempo, no modo como as diferentes organizações do território se relacionam entre si.

2.3 Similaridade entre grafos

Quando a informação está estruturada como um grafo, torna-se possível medir a similaridade entre dois termos desse grafo, bastando para isso calcular a similaridade dos subgrafos compostos por cada termo (“graph-matching”).

Gentleman (Gentleman, 2005) propõe que a similaridade entre grafos se defina pelo número de nós que são comuns aos dois grafos induzidos, divididos pelo número de nós contidos em pelo menos um dos dois grafos induzidos.

$$sim_{UI}(c_1, c_2) = \frac{|subGraph(root, c_1) \cap subGraph(root, c_2)|}{|subGraph(root, c_1) \cup subGraph(root, c_2)|}$$

Assim, quanto mais similar é c_1 de c_2 , maior é o seu subgrafo comum

2.4 Similaridade Semântica

De acordo com (Pesquita, 2009), o conceito de Similaridade Semântica, num dado domínio, avalia e determina o grau de semelhança entre dois conceitos. A identificação do grau de semelhança entre dois conceitos pressupõe a apreensão dos seus significados numa determinada ontologia, definindo, esta última, o relacionamento existente entre os termos e, por conseguinte, as relações pai/filhos do grafo.

2.4.1 Medidas baseadas no comprimento do caminho

Rada (Rada et al, 1989) demonstrou que os métodos simples se baseiam no caminho, sendo a distância, entre dois termos num caminho, calculada pela contagem do número de nós que se interpõem entre esses dois termos.

Foi proposta por Leacock e Chodorow (Leacock and Chodorow, 1998) a introdução de uma medida normalizada que usa o dobro da profundidade máxima na taxinomia com o objectivo de escalar o comprimento do caminho entre termos. Outra abordagem foi proposta por Wu e Palmer (Wu and Palmer, 1994), que integram a profundidade dos dois nós na análise e, simultaneamente, em relação a esses nós, a profundidade do antecessor comum mais baixo. Contudo, como demonstrou Budanitsky (Budanitsky, 1999), as duas abordagens supracitadas partem da premissa de que estamos perante uma distribuição uniforme dos nós e dos termos e, adicionalmente, de que os nós se encontram, na totalidade, ao mesmo nível, o que corresponderá a uma distância semântica equivalente. Assim, e para ultrapassar esta limitação do modelo, diversas abordagens adicionaram um peso diferenciado aos nós, permitindo distingui-los hierarquicamente quanto à sua profundidade implícita.

2.4.2 Medidas baseadas no Conteúdo de informação

A utilização do conteúdo de informação, variando a distância do nó que liga dois conceitos consoante a informação partilhada por estes (Resnik, 1995), induz um critério de ponderação que permite distinguir nós com semelhante profundidade na ontologia. A medida do Information Content (IC) de um termo resulta de um valor inversamente proporcional à sua frequência no “corpus”, indicando a sua especificidade. A maior especificidade de um termo, então, significa um valor superior do IC a ele associado (Couto, 2006).

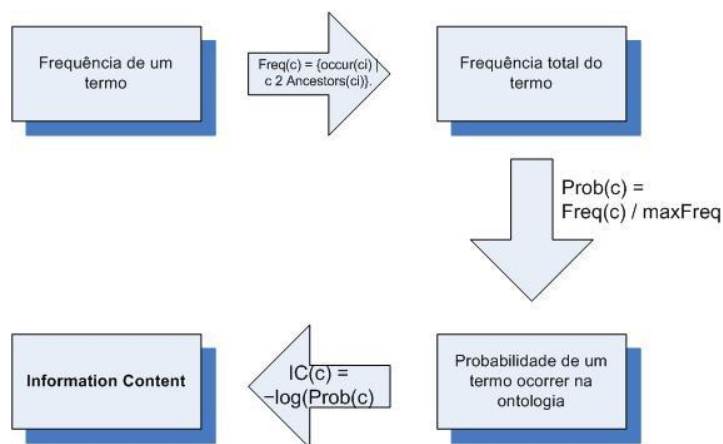


Figura 2.1: Esquema para o cálculo do IC

A frequência de um conceito ($Freq(c)$) pode ser definida pelo número de vezes que esse termo e todos os seus descendentes ocorrem num corpus. Isto significa que a raiz do grafo irá conter o valor referente à frequência máxima ($maxFreq$).

Uma maneira de observar a probabilidade de um conceito é dada pela fórmula

$$Prob(c) = \frac{Freq(c)}{maxFreq}$$

O valor do IC de cada conceito c pode ser encontrado através do valor negativo do logaritmo da probabilidade desse conceito c ocorrer

Assim, quanto maior for a probabilidade de encontrar um exemplo de um percurso possível até um termo no grafo (o qual define o conceito), menor será o seu IC, uma vez

que locais mais remotos, e por conseguinte menos encontrados no “corpus”, têm um IC mais elevado.

As medidas de semelhança semântica são medidas que se baseiam no uso do IC, calculando-se a informação partilhada entre dois termos, de modo a determinar a similaridade entre estes. Para este tipo de medidas, quanto maior for a informação partilhada por dois conceitos, maior é a semelhança semântica entre eles, podendo estas ser utilizadas para estabelecer a distância entre conceitos (Resnik, 1995).

Resnik (Resnik, 1995) propôs ainda uma abordagem em que a informação partilhada por dois conceitos é dada através do cálculo do IC do antecessor comum mais informativo (MICA)² da taxonomia. Contudo, com esta abordagem, se dois pares de conceitos tiverem o mesmo MICA, o resultado da sua semelhança semântica será exactamente o mesmo.

$$Sim_{Res}(c_1, c_2) = IC(c_{MICA})$$

Foi então proposto por Jiang e Conrath (Jiang and Conrath, 1997) a introdução de uma abordagem que combina o conceito do IC com o número de nós entre os termos.

$$dist_{JC} = IC(c_1) + IC(c_2) - 2 \times IC(c_{MICA})$$

Outra abordagem foi feita por Lin (Lin, 1998), que sugere que no cálculo da semelhança semântica se considere tanto o que há de comum quanto o que há de diferente entre os conceitos. Desta forma, Lin considera que a semelhança semântica entre conceitos aumenta ou diminui, respectivamente, consoante haja mais aspectos comuns ou diferentes entre aqueles.

$$sim_{Lin}(c_1, c_2) = \frac{2 \times IC(c_{MICA})}{IC(c_1) + IC(c_2)}$$

Desta forma, esta medida tem mais em conta o quão perto estão os termos do MICA do que quão específico é esse MICA.

² Do termo em inglês Most Informative Content Ancestor

$$dist_{Lin}(c_1, c_2) = 1 - sim_{Lin}(c_1, c_2) = \frac{IC(c_1) + IC(c_2) - 2 \times IC(c_{MICA})}{IC(c_1) + IC(c_2)}$$

As medidas de similaridade supracitadas obtiveram uma elevada importância na área da bioinformática (Pesquita, 2009) e foi dentro deste ramo que se desenvolveram novas formas de cálculo da semelhança semântica entre termos de uma ontologia, nomeadamente o GRASM e outras medidas combinadas (Pesquita, 2009), as quais serão apresentadas de seguida.

2.4.3 GRASM - Graph-Based Similarity Measure

A larga maioria das medidas que são utilizadas para calcular a semelhança semântica a partir do IC dos termos, contempla somente o IC do MICA entre dois termos, produzindo o mesmo resultado para qualquer par de termos que partilhem o mesmo MICA. Couto (Couto, 2006) introduziu uma nova forma de usar o IC dos termos, o “Graph-Based Similarity Measure” (GRASM).

A grande inovação do GRASM é a de utilizar, além do IC do MICA, a média de todos os antecessores disjuntivos do par de termos do qual se pretende saber a semelhança semântica, gerando-se, assim, um valor mais correcto e único para cada par de termos que tenham o mesmo MICA.

Esta nova abordagem pode ser aplicada a todas as medidas que usam o MICA, substituindo o IC do MICA por um valor que reflecta a média dos valores de IC de todos os antecessores disjuntivos.

2.4.4 Medida Híbrida

Tendo como base a medida Sim_{UI} desenvolvida por Gentleman’s, Pesquita (Pesquita, 2006) desenvolveu uma nova medida, a qual combina medidas de similaridade entre grafos e medidas baseadas no IC – a Sim_{GIC} Graph and Information Content Similarity.

A sim_{GIC} é uma medida que contabiliza a estrutura do grafo para calcular a similaridade. A informação contida nos nós é utilizada como um coeficiente de ponderação entre dois conceitos, possibilitando, deste modo, a produção de resultados mais precisos.

Trata-se de uma medida híbrida visto ter em conta o IC e, ao contrário de outras medidas de semelhança semântica que apenas utilizam o IC do MICA e dos termos dos quais se quer calcular a semelhança semântica, utiliza também o IC de todos os antecessores destes termos.

A semelhança semântica é calculada através do rácio entre a soma do IC dos termos que os dois subgrafos induzidos têm em comum e a soma do IC dos termos pertencentes à intersecção dos dois subgrafos. Este valor é dado através da seguinte fórmula:

$$sim_{GIC}(c_1, c_2) = \frac{\sum_{t \in t_{c_1} \cap t_{c_2}} IC(t)}{\sum_{t \in t_{c_1} \cup t_{c_2}} IC(t)}$$

em que t_{c_1} e t_{c_2} são termos dos subgrafos induzidos dos termos c_1 e c_2 respectivamente.

2.4.5 GKB - Geographic Knowledge Base

Tendo por base a definição de Grubbers para as ontologias, Chaves (Chaves et al., 2005) desenvolveu uma ontologia genérica para proceder ao tratamento de informação geográfica. Para Chaves, uma geo-ontologia é um conjunto de relações geográficas, definidas formalmente e sem qualquer ambiguidade. Este tipo de geo-ontologia pode ser usado por uma Geographic Knowledge Base (GKB).

A GKB Base constitui um repositório de informação geográfica baseada num domínio independente de meta-dados, capaz de integrar o conhecimento geográfico de diferentes fontes de dados, seguindo um conjunto de regras para a adição de informação no repositório e suportando a definição da relação da ontologia entre diferentes entidades em cada domínio. No domínio geográfico, providencia relações do tipo “parte de”, “adjacência”, e ainda outras relações entre entidades.

A GKB (Chaves, 2009) possui a capacidade de efectuar a distinção entre o nome e a característica (ou entidade) que a representa. As características e os seus nomes são classes diferentes e cada característica está associada a um tipo específico de característica.

Desta forma, a entidade “Concelho de Faro”, por exemplo, é associada ao tipo de entidade “concelho” e ao nome “Faro”. As entidades são classificadas por tipo de características ou propriedades. Esta distinção permite ao GKB estabelecer relações de “muitos para um” entre nomes e características, e ainda acrescentar novos tipos de dados.

Como demonstra a Figura 2.2, uma característica geográfica é composta por um nome (“name”) e tipo (“type”). A característica é associada ao Tipo (por exemplo, a entidade “Douro” tem o “tipo” associado de “Rio”), enquanto a classe Tipo de Relação capta a relação entre tipos e entidades (por exemplo “parte de”, “adjacente” a e outras formas de natureza geográfica). A classe Feature-Relationship obtém as relações entre entidades (por exemplo, o concelho de Sintra é parte de Distrito de Lisboa).

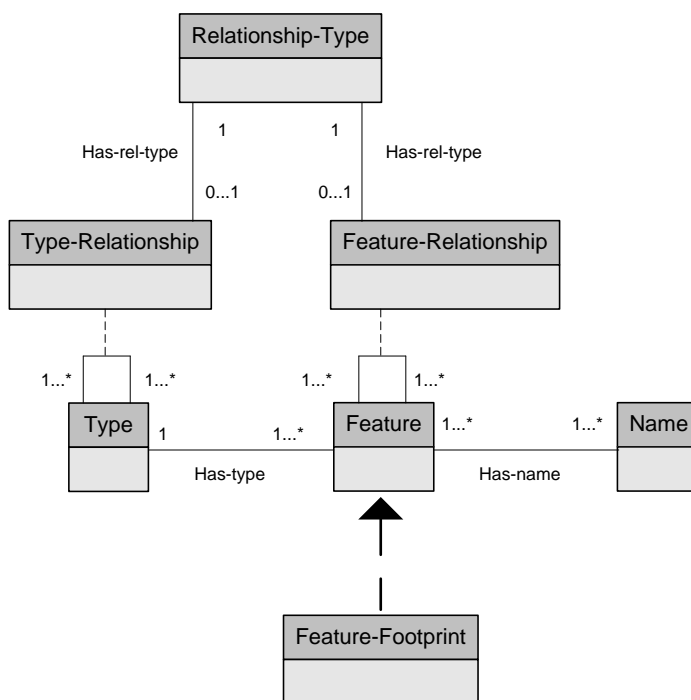


Figura 2.2: Meta-modelo do GKB. Fonte: Chaves et al., 2005

A partir desta construção de dados, Marcirio Chaves (Chaves, 2009) criou duas ontologias: a World Geographic Ontology (WGO), constituída por nomes geográficos de todo o Mundo, que foram obtidos através de dados publicados na Internet e a Geo-Ontologia de Portugal (Geo-Net-Pt), já referida anteriormente, e que será apresentada de forma mais detalhada em seguida.

2.4.6 Geo-Net-PT 01

A Geo-Net-PT 01, desenvolvida por Marcírio Chaves (Chaves et al, 2005), foi criada pelo grupo de pesquisa do XLDB, da Faculdade de Ciências da Universidade de Lisboa, no âmbito do projecto GREASE. Esta ontologia pode ser requisitada para fins científicos, sendo a Linguateca³, a entidade responsável pela sua disponibilização da mesma, através do seu Pólo no XLDB.

Enquanto ontologia, encontra-se estruturada de uma forma inteligível dentro dos formatos internacionalmente recomendados; evidencia as relações de “Parte de” (“part-of”), indicando que a “feature” em questão é, simultaneamente, parte do seu antecessor e “Adjacente”. Deste modo, duas referências geográficas situar-se-ão lado a lado, podendo as relações especificadas ser observadas na Figura 2.3.

A Geo-Net-PT 01 é constituída por dois domínios: o primeiro é de carácter geo-administrativo (contendo mais de 400.000 dados demográficos e administrativos de Portugal, incluindo informação, entre outros, sobre distritos, municípios, ruas e coordenadas geográficas); o segundo domínio abarca uma rede (contemplando informações relativas aos domínios “Web” e a “websites”).

No que diz respeito às instâncias geográficas desta ontologia, e uma vez que foi criada com base no GKB, a Geo-Net-PT 01 é composta por uma instância da classe “Feature” associando-se, a esta, um nome da classe “Name”. Para um conjunto de 266172 nomes na ontologia, existem 418744 “features”, revelando, à partida, um maior número de “features” relativamente ao número de nomes; como consequência, existem “features” diferenciadas que utilizam o mesmo nome, o que acarreta, naturalmente, e no momento da pesquisa do nome de um local, ambiguidade à associação estabelecida entre nomes e “features”. Desta forma, e como exemplo, o nome “Liberdade”, está relacionado com 486 “features”; a esse nome, na Geo-Net-PT 01, poderão corresponder

³ Linguateca – Centro de Recursos Distribuído para o processamento computacional da Língua Portuguesa, <http://www.linguateca.pt/>

15 conceitos diferentes na geografia portuguesa (tal como “rua”, “avenida” ou “praça”). Esta ambiguidade relativamente aos nomes pode ser observada na Tabela 2.1.

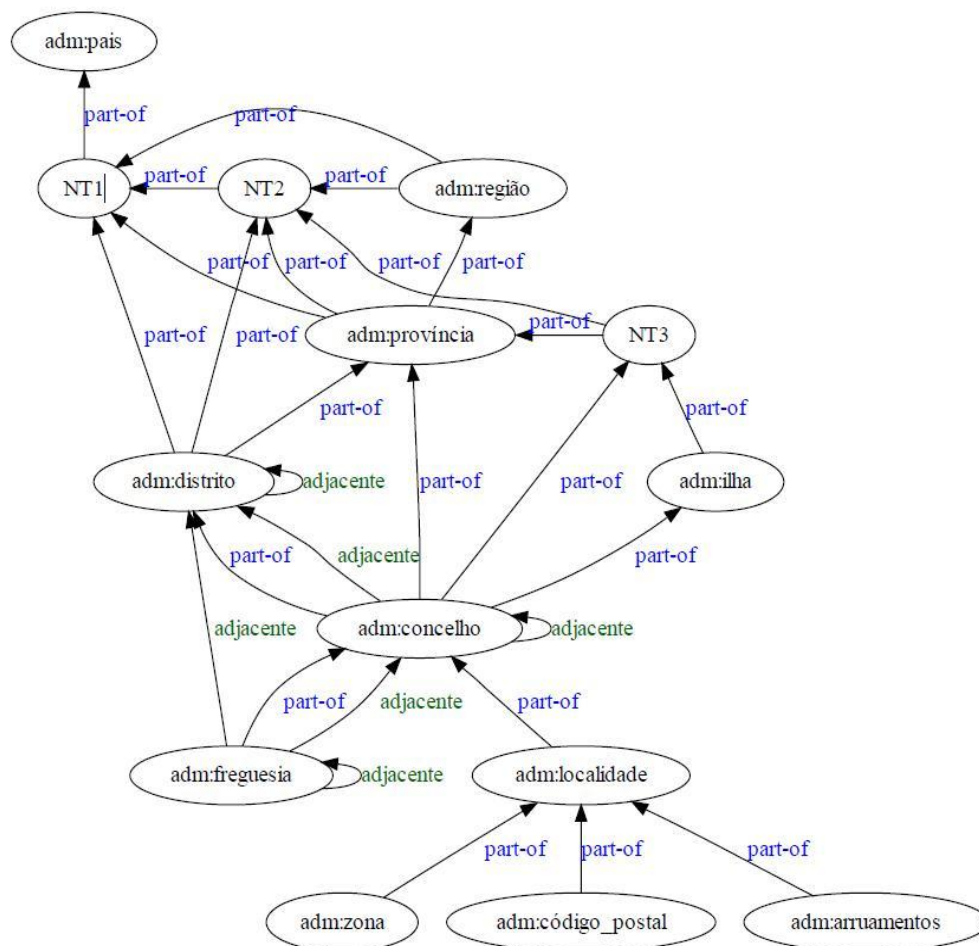


Figura 2.3: Relações entre tipos de conceitos para os dados físicos da Geo-Net-PT 01. Fonte: (Batista,2009).

# Palavras	Distinto	Todas as Palavras ambíguas (%)	≥1 palavra ambígua (%)
1	11.561	2.433 (21,04)	5.295 (45,78)
2	4.569	391 (8,34)	834 (18,25)
3	10.984	705 (6,42)	1.462 (13,31)
4	2.351	194 (8,25)	404 (17,18)
5	589	0	42 (7,13)
6	109	0	0
7	42	0	0
8	6	0	0
9	6	0	0
Σ	30.217	3.733 (12,35)	8.035 (26,59)

Tabela 2.1: Distribuição e ambiguidade dos termos na Geo-Net-PT por número de palavras.

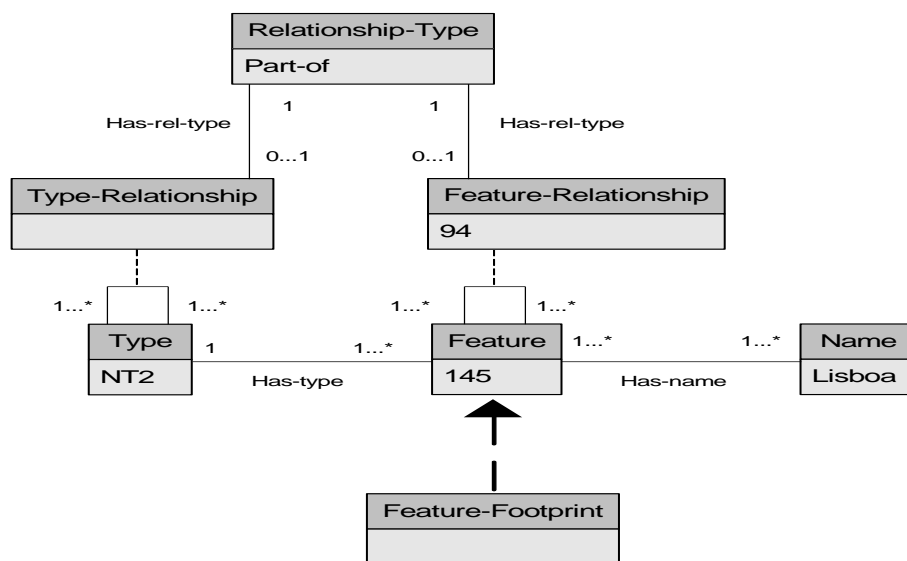


Figura 2.4: Exemplo de uma feature na Geo-Net-PT 01

Obtiveram-se os dados da Geo-Net-PT 01 a partir de diferentes tipos de fontes externas, sendo observáveis alguns erros que, devido à sua natureza, não são passíveis de correcção (Chaves et al, 2005). Deste modo, é possível verificar a existência de erros ortográficos, sabendo-se que a probabilidade do erro humano aumenta quando estamos perante a introdução de um conjunto alargado de dados. É frequente, ainda, surgirem inconsistências associadas à estrutura, atendendo a que a informação é organizada de forma diversa consoante a fonte utilizada. Quando estamos perante dados provenientes de fontes distintas, torna-se comum obter diversos nomes para o mesmo local, levando, este facto, à inclusão de nomes alternativos para uma mesma instância da classe “Feature”. Por exemplo, “São João”, localizado no Distrito de “Viana do Castelo”, possui “Vila Chã” e “São João Baptista” como nomes alternativos.

De modo a eliminar, parcialmente, as inconsistências referidas, é atribuído um nível de autoridade a cada fonte de informação, colmatando-se, por esta via, as inconsistências nos dados. (Chaves et al, 2005).

Numa tentativa de resolução de alguns dos problemas acima descritos, o grupo de pesquisa do GREASE decidiu efectuar uma limpeza aos dados da Geo-Net-PT 01. O resultado desta limpeza foi concretizado na Geo-Net-PT 02, uma ontologia que, apesar de se encontrar ainda em fase de implementação e em constante aperfeiçoamento, é disponibilizada através do “site” do GREASE II⁴.

⁴ http://xldb.di.fc.ul.pt/wiki/Geo-Net-PT_02

2.4.7 Geo-Net-PT 02

Em primeiro lugar, e como foi anteriormente aludido, esta ontologia é uma versão melhorada da sua antecessora Geo-Net-PT 01. Para isso contribuiu a solução encontrada quanto aos problemas associados ao conjunto de nomes e termos; além disso, foram acrescentadas novas características à ontologia melhorada.

O desenvolvimento da Geo-Net-PT 02 (versão 1) iniciou-se, por conseguinte, com o enriquecimento de informação da sua antecessora; adicionaram-se, por exemplo, dados administrativos como informação sobre concelhos, ruas, distritos, etc. e, a estes, acrescentaram-se dados do domínio físico (Viegas, 2006). Alguma da diferença de informação pode ser observada na Tabela 2.2.

A Geo-Net-PT 02 (versão 2) é uma versão melhorada da anterior; neste caso, foram acrescentados, entre outros, um novo conjunto de tipos de relação entre as características “has part”, “is located on”, “is part of”, etc. (Lopez et al., 2009).

No que diz respeito ao desenvolvimento do projecto analisado nesta tese, utilizou-se a ontologia Geo-Net-PT 02 na sua primeira versão; tal ficou a dever-se, no exacto momento de implementação do presente projecto, à indisponibilidade da versão mais actualizada da ontologia. É de referir que a ontologia utilizada foi instalada sob uma plataforma PostGresSQL⁵ que é um Sistema de Gestão de Bases de Dados relacional.

<i>Nome da Tabela</i>	<i>Geo-Net-PT 2.1</i>	<i>Geo-Net-PT 2.2</i>
adm_feature	418744	388049
adm_name	266172	265044
net_name	115588	23666
adm_feature_relationship	420237	423836
adm_feature_type	2	3
phy_feature	191	5662
phy_feature_attribute	0	8305
phy_feature_footprint	0	3208
phy_feature_relationship	0	2794
phy_name	105	8250
phy_name_attribute	0	4093

Tabela 2.2: Diferença nos dados entre a Geo-Net-PT 2.1 (versão usada) e a Geo-Net-PT 2.2.

⁵ <http://www.postgresql.org/>

2.5 Web 1T 5-gram Corpus (Google N-Grams)

O Web 1T 5-gram corpus, conhecido por Google N-Grams⁶, é um conjunto de dados disponibilizado pela Google Inc, criado para responder a vários tipos de pesquisa científica nomeadamente estatísticas de tradução automática, reconhecimento de fala ou correctores ortográficos automáticos.

Este conjunto de dados foi gerado a partir de aproximadamente 1 trilião de “tokens” de palavras retiradas de páginas públicas de publicidade e inseridas nas bases de dados da Google Inc.; Esta coleção de dados contem n-gramas de palavras inglesas, incluindo o nome de locais portugueses e a sua frequência observável no “corpus”. A distância dos n-gramas varia de um até cinco gramas. Aos 1-gramas estão associados nomes constituídos por apenas uma palavra e, no outro extremo, os 5-gramas serão nomes constituídos por 5 palavras.

Apesar do vasto leque de nomes que se podem encontrar no “corpus” do Google N-Grams, este conjunto de dados não inclui todos os nomes que constam na Geo-Net-PT 02.

A Tabela 2.3 ajuda-nos a perceber melhor este grau de inclusão. No Google N-Grams estão representados 84,205% dos nomes existentes na Geo-Net-PT 02; assim, e em termos representatividade de instâncias geográficas, somente 52,741% se encontram presentes na ontologia.

De modo a ter um acesso automatizado aos dados do Google N-Gram’s, utilizou-se o jWeb1T, uma ferramenta desenvolvida em JAVA de código livre sobre o qual analisaremos no ponto seguinte.

	Geo-Net-Pt	Found in GNGC	Percentage
Unique Names	78070	65739	84.205%
Feature Names	199053	104982	52.741%

Tabela 2.3:Quantificação dos nomes da Geo-Net-PT 02 contidos no Google N-Grams Corpus.

⁶ <http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html>

2.6 jWeb1T

O jWeb1T [web1t], como aludimos acima, é uma ferramenta de código livre, disponível através do site <http://tcc.itc.it/research/textec/tools-resources/jweb1t/user-guide-1.0.html#Introduction>. Desenvolvida em JAVA, utiliza-se para efectuar pesquisas eficientes no “corpus” disponibilizado pela Web 1T 5-gram corpus. O seu método de pesquisa baseia-se no algoritmo de pesquisa binário, dependendo a pesquisa de um n-grama num arquivo específico do número de palavras que a frase contém; por fim, a pesquisa devolve a contagem da sua frequência dentro de um tempo logarítmico.

O “corpus” encontra-se guardado em diversos ficheiros, sendo por isso utilizado um índice simples, de modo a devolver os ficheiros que contêm os n-gramas começados com um prefixo específico. Para que o jWeb1T consiga aceder ao “corpus” Google N-Grams, torna-se imprescindível que este esteja instalado e descomprimido no disco do computador.

Capítulo 3 GeoSSM

O primeiro objectivo deste projecto visava a obtenção de valores de semelhança semântica entre termos de uma ontologia geográfica. Para tal, foi desenvolvida a GeoSSM, apresentada neste capítulo. São descritos, pormenorizadamente, os métodos implementados, assim como o procedimento utilizado para o cálculo das medidas de semelhança semântica entre dois termos.

3.1 Arquitectura

A GeoSSM é uma ferramenta criada em Java que calcula a semelhança semântica entre quaisquer dois termos da Geo-Net-PT a partir dos seus ID's, bastando, para isso, eleger entre as diferentes medidas de semelhança semântica, aquela com que se quer obter o resultado (Ver Figura 3.1).

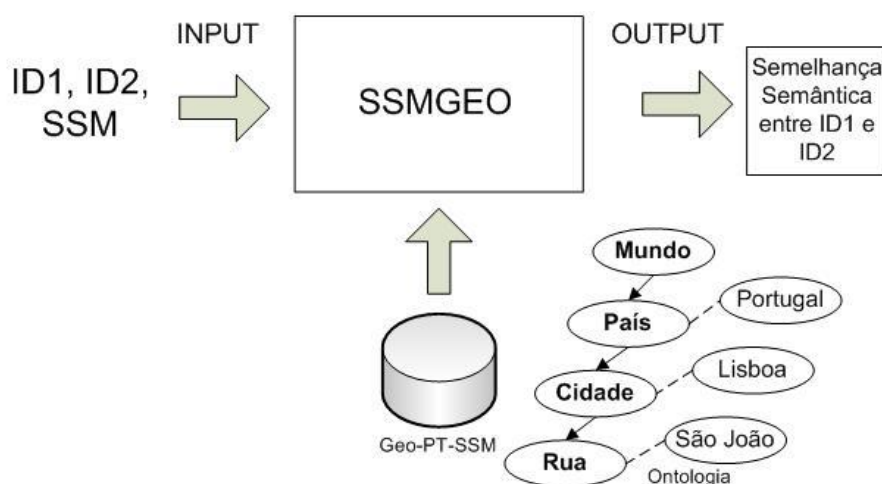


Figura 3.1: Arquitectura da GeoSSM.

A GeoSSM implementa as medidas de semelhança semântica tal como descritas na secção 2.4 (Similaridade Semântica, página 20). Para tal foi adaptado o código do ProteinOn (Pesquita, 2006) utilizando a Geo Net PT como estrutura ontológica e o “corpus” do Google N-Grams para o cálculo do IC dos termos. Foram ainda introduzidas algumas alterações à Geo-Net-PT, procurando-se, assim, melhorar a sua adaptação ao processo de cálculo da semelhança semântica entre os seus termos.

3.2 Alterações à Geo-Net-PT

De modo a ser possível efectuar cálculos de semelhança semântica entre dois termos, adicionaram-se duas novas tabelas à Geo-Net-PT 02; de facto, esta, por si só, não fornecia toda a informação necessária à implementação das medidas de semelhança semântica descritas nos capítulos anteriores. Foi então criada, a partir da Geo-Net-PT 02, uma nova base de dados – a Geo-Net-PT-SSM -, instalada no servidor Agatha, pertencente ao grupo do XLDB. Uma vez que esta nova ontologia é uma extensão da Geo-Net-PT 02, foi igualmente gerada sobre a tecnologia PSQL.

As tabelas acrescentadas à base de dados dão pelo nome de “ssm_graphpath” e “ssm_termfreq”.

Para se conservar a coerência de nomes usados na estrutura da ontologia, a criação destas novas tabelas obedeceu à preocupação de manter o nome dos identificadores, dos termos das tabelas e das relações entre esses termos da Geo-Net-PT 02.

Assim sendo, uma instância da tabela “ssm_termfreq” será identificada pela coluna “f_id”, tal como na tabela “adm_feature” da Geo-Net-PT 02.

A tabela “ssm_graphpath”, por sua vez e tal como acontecia com a tabela “adm_feature_relationship” na Geo-Net-PT 02, indica relações entre dois termos; no entanto, a “adm_feature_relationship” apresentava somente relações directas, sendo que o termo identificado em “f_id2” é “part-of” (sucessor) ou “adjancy” do termo identificado em “f_id1”. A “ssm_graphpath”, por seu turno, demonstra relações tanto directas como indirectas de “part of” entre dois termos. Portanto, (“f_id2” identifica um termo que pode ser sucessor directo ou indirecto do termo identificado em “f_id1”).

3.2.1 Tabela SSM_GraphPath

Como foi referido acima, esta tabela descreve todas as relações existentes entre um termo “f_id2” e todos os seus antecessores “f_id1,” quer sejam antecessores directos ou indirectos, até ao topo do grafo (Tabela 3.1).

Column	Type	Modifiers
graph_path_id	Integer	not null default 0
f_id1	Integer	not null default 0
f_id2	Integer	not null default 0
Distance	Integer	not null default 0
Indexes:	"graphpath_key" PRIMARY KEY, btree (graph_path_id)	
	"ssm_graphpath_f_id1_key" UNIQUE, btree (f_id1, f_id2, distance)	

Tabela 3.1:Esquema da Tabela "ssm_graphpath" da base de dados Geo-Net-PT-SS

A tabela “ssm_graphpath” associa ainda um valor a cada ligação (Distance), que indica a distância a que se encontram os termos em questão, a distância que é calculada pelo número de ligações existentes entre dois termos.

As Tabelas e Figuras apresentadas abaixo, representam o Subgrafo do termo que corresponde ao “Concelho de Lisboa” cujo ID = 146 na Geo-Net-PT-SSM, sendo que a Tabela 3.2 e Figura 3.2 são representativas do subgrafo gerado pela tabela “adm_feature_relationship”, e a Tabela 3.3 e Figura 3.3 são representativas do subgrafo gerado pela tabela “ssm_graphpath”.

<i>fr_id</i>	<i>f_id1</i>	<i>f_id2</i>	<i>frt_id</i>	<i>is_id</i>
206	129	146	PRT	100
196699	3965	146	PRT	200
420258	418732	146	PRT	501

Tabela 3.2: Relação de “part-of” entre o termo 146 (Lisboa) e os seus antecessores directos na tabela “adm_feature_relationship” (SELECT * FROM adm_feature_relationship WHERE f_id2 = 146 and frt_id = 'PRT';)

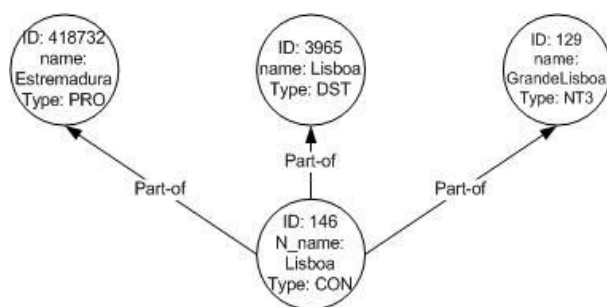


Figura 3.2: Subgrafo que correspondente à relação de “part-of” entre o termo 146 (Lisboa) e os seus antecessores que pode ser observada na tabela “adm_feature_relationship”.

<i>graph_path_id</i>	<i>f_id1</i>	<i>f_id2</i>	<i>distance</i>
5072725	418732	146	1
2681011	3965	146	1
887558	129	146	1
2681016	94	146	2
887562	145	146	2
2681021	418745	146	3
887568	94	146	3
887572	418745	146	4

Tabela 3.3: Relação de “part-of” entre o termo 146 (Lisboa) e os seus antecessores directos e indirectos na tabela “ssm_graphpath” (select * from ssm_graphpath where f_id2= 146 order by distance).

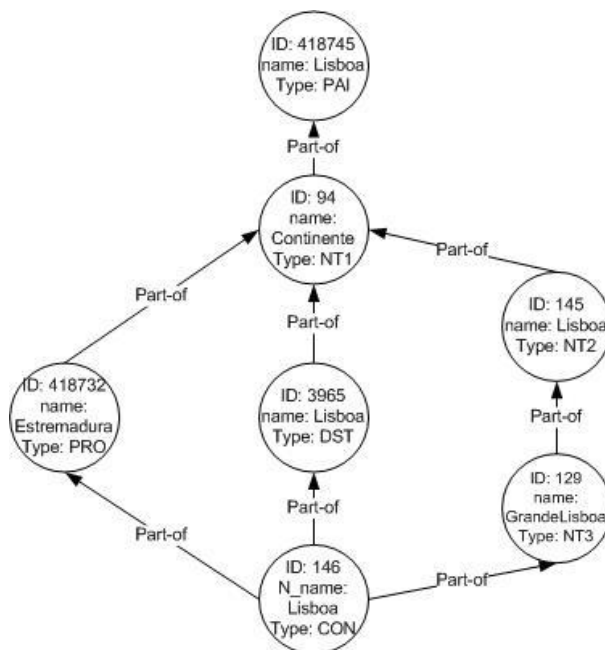


Figura 3.3: Subgrafo que correspondente à relação de “part-of” entre o termo 146 (Lisboa) e os seus antecessores que pode ser observada na tabela “ssm_graphpath”.

Comparando as tabelas e figuras acima referidas, pode-se constatar que a “ssm_graphpath” torna o cálculo da semelhança semântica mais simples e rápido. Dado que, para o cálculo da semelhança semântica é necessário ter uma visão total do subgrafo criado por cada um dos termos e os seus antecessores até à raiz da ontologia, a “ssm_graphpath” permite ter esta visão através de apenas com uma “query”. Já recorrendo à “adm_feature_relationship” seriam necessárias 6 “queries” para obter a mesma visão.

3.2.2 Tabela SSM_TermFreq

A tabela “ssm_termfreq”, resumida na Tabela 3.4, guarda os valores de IC de cada termo da Geo-Net-PT-SSM, bem como todos os valores necessários ao seu cálculo, os quais serão posteriormente utilizados pelas medidas que recorrem ao IC para medir a semelhança semântica entre termos.

Column	Type	Modifiers
f_id	Integer	not null
freq	double precision	not null default 0::double precision
hfreq	double precision	not null default 0::double precision
prob	double precision	not null default 0::double precision
info_content	double precision	not null default 0::double precision
rel_info	double precision	not null default 0::double precision
Indexes : "ssm_termfreq_key" PRIMARY KEY, btree (f_id)		

Tabela 3.4 Tabela " ssm_termfreq " da base de dados Geo-Net-PT-SSM.

Os campos desta tabela são os seguintes:

- **f_id**
Guarda as chaves primárias da tabela. Este campo é uma chave estrangeira da tabela “adm_feature”.
- **freq**
Esta coluna guarda as frequências do nome de cada termo da tabela “adm_feature” da Geo.
Este valor equivale à frequência com que um nome ocorre no “corpus” do Google-Ngrams. A frequência de cada termo equivale à soma da

frequência encontrada para as diferentes formas de representar um nome na Geo-Net-PT-SSM (caracteres “simple ASCII”, “capitalized” (letras maiúsculas), e “non-capitalized” (Primeira letra do nome maiúscula e as restantes minúsculas)). Ex: “são João”, “são João” e “São João”.

A ontologia Geo-Net-PT 02 contém locais, cujos nomes não foram encontrados no Google N-Grams. Para estes casos, foi-lhes atribuído o valor 1. Esta abordagem permite que não haja a possibilidade de calcularmos o logaritmo de 0 no cálculo do IC, o que é necessário no uso de algumas medidas de semelhança semântica, como é o caso do algoritmo de Resnik (Sim_{Resnik}) visto ser uma medida que devolve directamente os valores de IC, nomeadamente o IC do MICA associado ao par de termos que estamos a comparar.

- **hfreq**

Corresponde à frequência de cada termo, previamente calculado em “freq”, e à soma das frequências de todo os seus sucessores (filhos).

- **prob**

Guarda o valor que indica qual a probabilidade de ocorrência de um termo numa pesquisa geográfica.

Este valor é obtido através da fórmula:

$$prob(c) = \frac{Freq(c)}{maxFreq}$$

- **info_content**

Guarda o valor do IC que cada termo possui.

$$IC(c) = -\log(Prob(c))$$

- **rel_info**

Guarda os valores normalizados do IC. Este campo aglomera o conjunto de dados de maior importância da tabela, uma vez que os valores aqui guardados serão usados pelas medidas de semelhança semântica que

utilizam o IC para o cálculo da semelhança semântica entre dois termos da ontologia.

Os valores desta coluna podem variar entre 0 e 1. Um termo com um IC relativo próximo de 0 será um valor pouco ou nada informativo, isto é, será o termo mais geral do grafo que corresponde à sua raiz. Em contrapartida, os termos com um IC relativo de 1 serão os termos mais informativos e equivalerão aos termos cujos nomes têm pouca frequência ou não foram encontrados no “corpus” do Google N-Grams. A não inclusão desses nomes no vasto leque de palavras usadas na publicidade do Google dever-se-á, muito provavelmente, ao facto de estarem em causa referências geográficas pouco frequentes na Internet.

O facto dos valores da “rel_info” estarem normalizados entre 0 e 1, dá-nos uma melhor percepção do quão informativo é o IC de um termo; da sua análise conseguimos obter um melhor entendimento do seu valor em relação ao valor dos restantes termos, sabendo-se, à partida, qual o valor máximo e o valor mínimo existentes no grafo.

$$IC_{Rel}(c) = \frac{-\log_2 f(c)}{\log_2 A}$$

em que A são todos os antecessores de c, tanto directos como indirectos.

3.2.3 Estrutura

Como já foi referido anteriormente, as medidas de semelhança semântica podem ser aplicadas a estruturas de dados organizados como um DAG.

A Geo-Net-PT-SSM pode ter esta organização estrutural se ignorarmos, entre os termos, as relações de “adjacency” da tabela “adm_feature_relationship”. Com esta visão do grafo, é possível, sem que haja a possibilidade de ocorrerem ciclos na sua procura, efectuar as inferências necessárias aos cálculos da semelhança semântica entre termos, encontrando-se, por exemplo, os ancestrais comuns mais informativos de dois termos ou ainda o subgrafo que lhes diga respeito.

O DAG da Geo-Net-PT-SSM está então assente nas relações de “part-of” entre termos geográficos e o IC destes. Por isso, interessa para o cálculo da semelhança semântica, averiguar se os termos são parte de um mesmo “Concelho”, de uma mesma “Zona” ou de uma qualquer outra divisão territorial, atendendo às várias formas de organização do território nacional - nomeadamente NUTS (I,II e III) ⁷, Províncias ou Freguesias. Estas relações de “part-of” podem ser observadas na Figura 3.4. Constituindo uma característica fundamental num DAG, constata-se que, partindo de qualquer termo da ontologia, é impossível estabelecer um caminho de retorno ao ponto de partida

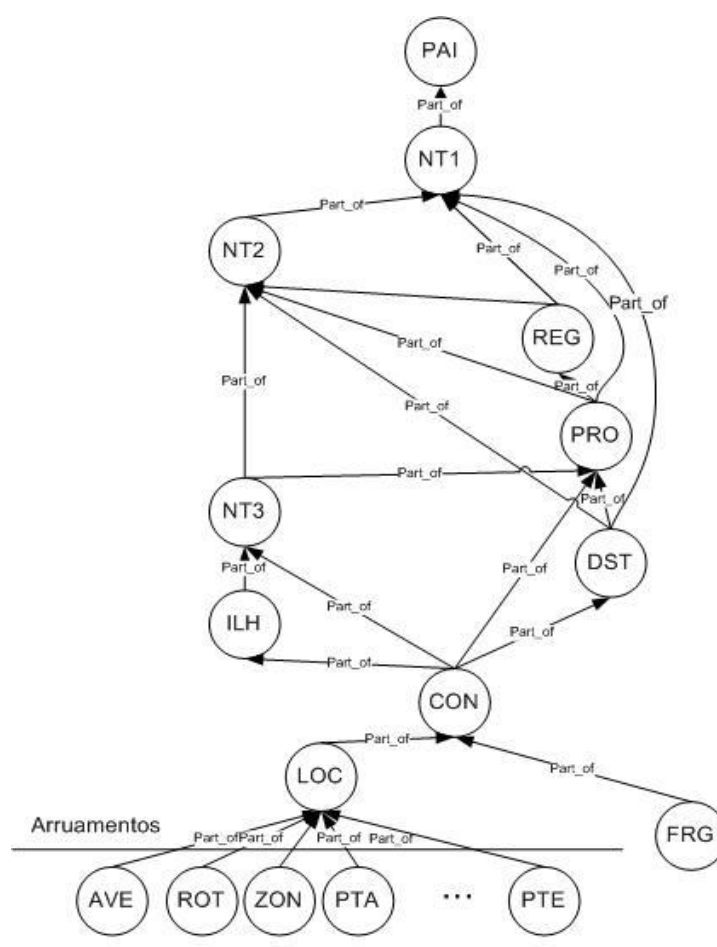


Figura 3.4: Relações do tipo “Part of” para os dados físicos da Geo-Net-PT.

⁷ NUTS- Nomenclatura das Unidades Territoriais para fins estatísticos. Trata-se de uma divisão e classificação do território nacional em regiões estatísticas de acordo com o estabelecido para a União Europeia.

De seguida, apresentar-se-ão alguns exemplos do cálculo de semelhança semântica entre termos da Geo-Net-Pt-SSM produzidos pela GeoSSM, dando-se ênfase às limitações das medidas implementadas e procurando demonstrar-se que estas estão presentes na ontologia utilizada.

3.3 Exemplos

Nesta secção, foram utilizados conjuntos de termos com idêntico valor de IC (uma vez que existem bastantes termos nestas condições, como se pode verificar na Tabela 3.5), conjuntos de termos cujos pares têm o mesmo MICA e, ainda, conjuntos de termos com idênticos antecessores comuns e cujos pares, constituídos por esses termos, possuem o mesmo MICA.

<i>term_freq features</i>	<i>features (IC igual)</i>	<i>Percent. (%)</i>
199053	104981	52.74

Tabela 3.5: Percentagem de features com o mesmo valor de IC na Geo-Net-PT.

3.3.1 1º Exemplo de Medidas de Semelhança Semântica (MICA e número de caminhos igual)

Foram aqui considerados os termos “182631” (Gago Coutinho), “197748”(Padre Cruz) e “224027”(25 de Abril), cujo valor de IC e todos antecessores de cada termo podem ser observados na Tabela 3.6

<i>Termo ID</i>	<i>Nome</i>	<i>IC</i>	<i>Ancestors</i>
182631	Gago Coutinho	0.24151903	[182400, 31, 129, 3965, 418732, 145, 94, 418745]
197748	Padre Cruz	0.25370184	[196518, 146, 129, 3965, 418732, 145, 94, 418745]
224027	25 de Abril	0.2237079	[223997, 198, 129, 3965, 418732, 145, 94, 418745]

Tabela 3.6: Termos utilizados para verificar o resultado das medidas de semelhança semântica com pares de termos cujo MICA é comum, e cujos termos estão localizados na ontologia à mesma distancia do MICA.

Como é possível verificar na Tabela 3.7 e na Figura 3.5, o número de nós “paths” existentes entre cada termo usado é igual (número de “paths” = 6), e o MICA (ID:129, Grande Lisboa, IC: 0.056471765) é partilhado por todos os pares.

$SSM(t1, t2)$	(182631,197748)	(182631, 224027)	(197748,224027)
Paths(#)	6	6	6
MICA(id)	129	129	129
MICA(ic)	0.056471765	0.056471765	0.056471765
CommonDisjAnc(#)	1	1	1
CommonDisjAnc	[129]	[129]	[129]
EdgeNaive	0.142857143	0.142857143	0.142857143
UI	0.428571429	0.428571429	0.428571429
Resnik	0.001516997	0.001516997	0.001516997
ResnikGrasm	0.007059796	0.007059796	0.007059796
Lin	0.228066984	0.242770825	0.236575682
LinGrasm	1.061377484	1.129806179	1.100975238
JiangConrath	0.004739795	0.006034387	0.005486935
JiangConrathGrasm	0.004875254	0.006189916	0.005486935
GIC	0.358510585	0.343672639	0.361087629

Tabela 3.7: Valores de semelhança semântica entre os termos 182631, 197748 e 224027 da Geo-Net-PT-SSM nas diversas medidas de semelhança semântica.

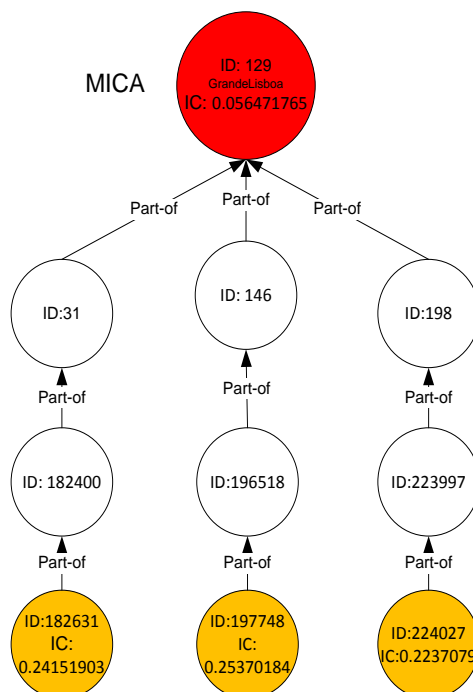


Figura 3.5: Subgrafo exemplificativo dos “paths” entre os termos 182631, 197748 e 224027 e o termo 129 (MICA comum a todos os pares constituídos por esses termos).

Ainda relativamente à Tabela 3.7 denote-se que, como esperado, nalgumas medidas não se produziu uma diferenciação no cálculo de semelhança semântica entre os conjuntos de pares constituídos por termos diferentes. Foi o caso da EdgeNaive , e da Sim_{UI} que tomaram em consideração apenas o número de caminhos entre um termo e outro e também o caso da Sim_{Resnik} por ter em conta apenas o valor do MICA.

3.3.2 2º Exemplo de Medidas de Semelhança Semântica (MICA e IC igual, sem antecessores disjuntivos)

Foram escolhidos os termos “418454” (Encarnação), “3646” (Grijó) e “2178”(Amareleja) pois têm IC de valor idêntico e 0 termos disjuntivos

Como podemos constatar na Tabela 3.8 e Tabela 3.9, a maioria das medidas conseguiram produzir, no exemplo anterior e entre pares de termos diferentes, valores de semelhança semântica diferenciado; no entanto, neste segundo exemplo, não foi possível obter os mesmos resultados uma vez que os termos utilizados, para além de partilharem o mesmo MICA, apresentam um valor de IC semelhante.

Acresce, ainda, que mesmo substituindo o valor de IC pelo GRASM para o cálculo da semelhança semântica, não é possível conseguir essa diferenciação, pois os pares de termos escolhidos apresentam os mesmos antecessores comuns (ID: 94 e ID: 418745), não havendo por isso antecessores disjuntivos entre os termos de cada par. A única exceção, neste exemplo, foi a medida Sim_{GIC}, que consegue essa diferenciação por ter também em consideração o valor de IC de todos os termos que estão entre os termos do par e o MICA.

<i>Termo ID</i>	<i>Nome</i>	<i>IC</i>	<i>Ancestors</i>
418454	Encarnação	1	[146, 129, 3965, 418732, 145, 94, 418745]
3646	Grijó	1	[331, 130, 3967, 418739, 418736, 196, 94, 418745]
2178	Amareleja	1	[186, 48, 3945, 12, 418734, 12, 94, 418745]

Tabela 3.8: Termos utilizados na 2º exemplo das medidas de semelhança semântica (Termos sem antecessores disjuntivos).

$SSM(t_1, t_2)$	(418454, 3646)	(418454, 2178)	(3646, 2178)
Paths(#)	7	7	8
MICA(id)	94	94	94
MICA(ic)	0.026511809	0.026511809	0.026511809
EdgeNaive	0.125	0.125	0.111111111
UI	0.111111111	0.117647059	0.105263158
Resnik	7.12E-04	7.12E-04	7.12E-04
ResnikGrasm	0.001402528	0.001402528	0.001402528
Lin	0	0	0
LinGrasm	0	0	0
JiangConrath	9.48E-06	9.48E-06	9.48E-06
JiangConrathGrasm	1.86E-05	1.86E-05	9.48E-06
GIC	0.066816448	0.058550128	0.056070111

Tabela 3.9: Valores de semelhança semântica entre os termos 418454, 3646 e 2178 da Geo-Net-PT nas diversas medidas de semelhança semântica.

3.3.3 3º Exemplo de Medidas de Semelhança Semântica (MICA e antecessores comuns iguais)

A grande vantagem da medida híbrida, face às que usam o MICA e o IC dos termos, é levar em consideração a intersecção e a união dos subgrafos produzidos entre cada termo até à raiz do grafo (factor de diferenciação como referido na secção 2.4.4); assim, procurou apurar-se o valor de semelhança semântica calculado pelas medidas entre pares de termos que partilham os mesmos antecessores.

Os termos escolhidos para este exemplo foram “1704”(Ajuda), “1705”(Alcântara) e “1709”(Benfica).

ID	Nome	IC	Ancestors
1704	Ajuda	0.189557	[146, 129, 3965, 418732, 94, 145, 94, 418745, 418745]
1705	Alcântara	0.212383	[146, 129, 3965, 418732, 94, 145, 94, 418745, 418745]
1709	Benfica	0.185193	[146, 129, 3965, 418732, 94, 145, 94, 418745, 418745]

Tabela 3.10: Termos utilizados no 3º exemplo de medidas de semelhança semântica (Termos com subgrafo de antecessores iguais).

$SSM(t1, t2)$	(1704, 1705)	(1704, 1709)	(1705, 1709)
Paths(#)	2	2	2
MICA(id)	146	146	146
MICA(ic)	0.08206718	0.08206718	0.08206718
EdgeNaive	0.333333333	0.333333333	0.333333333
UI	0.636363636	0.636363636	0.636363636
Resnik	0.002204565	0.002204565	0.002204565
ResnikGrasm	0.00926436	0.00926436	0.00926436
Lin	0.40835516	0.437982567	0.412837397
LinGrasm	1.716052885	1.840557732	1.73488883
JiangConrath	0.009504803	0.011393917	0.0097884
JiangConrathGrasm	0.009780935	0.01171773	0.0097884
GIC	0.767589677	0.767589677	0.767589677

Tabela 3.11: Valores de semelhança semântica entre os termos 1704, 1705 e 1709 da Geo-Net-PT nas diversas medidas de semelhança semântica.

Ao observarmos os valores apresentados na Tabela 3.11, verificamos que no cálculo da semelhança semântica entre pares de termos que partilhem o mesmo MICA e também os mesmos antecessores, a medida Sim_{GIC} produz valores semelhantes para os diferentes pares. Já as medidas Sim_{Lin} e $Sim_{JiangConrath}$ obtiveram diferenciação nos resultados pois neste exemplo foram usados termos com IC diferente (Tabela 3.10).

3.4 Medida Combinada - Sim_{Geo}

Com o objectivo de colmatar as limitações das medidas de semelhança semântica referidas anteriormente (secção 2.4) exemplificadas na secção 3.3, sugeri uma forma alternativa de produzir valores de semelhança semântica.

Esta necessidade justifica-se pelo facto de na Geo-Net-PT-SSM encontrarmos inúmeros termos com o mesmo valor de IC e de não existirem muitas organizações diferentes do território de Portugal, o que, conseqüentemente, poucos termos disjuntivos entre quaisquer dois termos. No cálculo de semelhança semântica entre termos da “Gene Ontology” (Pesquita, 2006), foram obtidos bons resultados com as medidas Sim_{GIC} e as medidas que usam o valor do GRASM em alternativa ao valor de IC, exactamente porque as condições indicadas acima não se verificavam

Uma sugestão para contornar as limitações acima descritas, far-se-ia através da utilização, por um lado, das vantagens das medidas que utilizam a partilha de informação entre os termos do par e o MICA (como acontece com a $Sim_{JiangConrath}$) e, por outro, através do benefício associado às vantagens do Sim_{GIC} . Assim, o valor de semelhança semântica seria calculado a partir da média das duas medidas enunciadas, criando-se e implementando-se a Sim_{Geo} .

$$Sim_{Geo} = \frac{(Sim_{GIC} + Sim_{JiangConrath})}{2}$$

3.4.1 Exemplo de Medidas de Semelhança Semântica com a Sim_{Geo}

Nesta secção realiza-se o cálculo de semelhança semântica entre os termos utilizados nas demonstrações anteriores (Secção 3.3 Tabela 3.12), aplicando a Sim_{Geo} com o objectivo de evidenciar que esta proposta consegue produzir resultados diferenciados para os diferentes cenários representados.

	<i>t1</i>	<i>t2</i>	<i>t3</i>
1º Exemplo	182631” (Gago Coutinho)	“197748”(Padre Cruz)	“224027”(25 de Abril)
2º Exemplo	418454” (Encarnação)	“3646” (Grijó)	“2178”(Amareleja)
3º Exemplo	“1704”(Ajuda)	“1705”(Alcântara)	“1709”(Benfica)

Tabela 3.12: Termos utilizados em cada exemplo no cálculo da semelhança semântica.

	$SSM(t1,t2)$	$SSM(t1,t3)$	$SSM(t2,t3)$
1º Cálculo	0.18822049	0.181440233	0.18987763
2º Cálculo	0.395110963	0.396043002	0.395250882
3º Cálculo	0.04003961	0.03590645	0.034666441

Tabela 3.13: Valores de semelhança semântica entre os termos usados em cada um dos exemplos anteriores aplicados à nova medida implementada Sim_{Geo} .

Através da observação dos resultados da Tabela 3.13, pode-se constatar que a Sim_{Geo} ultrapassa, à primeira vista, as limitações que as medidas anteriormente descritas apresentavam, de facto, a nova medida não produziu nenhum valor idêntico de semelhança semântica entre os diferentes pares de termos utilizados.

Capítulo 4 GeoScope - Geographical

Scope

No capítulo anterior descreveu-se a implementação das medidas de semelhança aplicadas a termos geográficos através da Geo-Net-PT-SSM. , que foram usadas neste projecto, também para desambiguar nomes de locais de Portugal presentes nos resumos de referências e atribuir o âmbito geográfico a esses mesmos resumos.

Dado que não é possível calcular o âmbito geográfico de duas referências geográficas sem conhecer exactamente a que locais nos estamos a referir (no fundo, quais os termos da Geo-Net-PT-SSM que as representam), e para dar resposta ao objectivo acima referido, foi criada, ao longo deste projecto, uma ferramenta em linguagem Java, o GeoScope. Esta ferramenta atribui o âmbito geográfico a um conjunto de referências geográficas, encontrando, por seu turno, a referência ontológica que, com maior probabilidade, melhor representa cada uma das referências geográficas por desambiguar.

De entrada, e entre um conjunto de estratégias implementadas na própria ferramenta, o GeoScope recebe o nome da estratégia que se quer utilizar para a desambiguação das referências geográficas, o nome da medida de semelhança e um conjunto de referências geográficas. A partir daqui, o GeoScope irá desambiguar as referências geográficas e definir o âmbito geográfico e ontológico que melhor se adequa às referências geográficas por desambiguar, recorrendo, para isso, às medidas de semelhança semântica implementadas no sistema GeoSSM.

4.1 Desambiguação de Referências Geográficas.

Para proceder à desambiguação das referências geográficas, são tidos em conta os termos da ontologia com esses nomes num todo, utilizando as referências geográficas que não são ambíguas (ou as que são menos ambíguas), de modo a clarificar as restantes.

Borges (Borges et al, 2007) constata que, numa pesquisa geográfica, quando um utilizador pretende referir um local específico, i.e, o caso de um arruamento (“rotunda”, “praça”, “rua”, “praceta”) inclui essa denominação, por exemplo, “... na Avenida da Liberdade”, “... na Rua Gago Coutinho”. Já no caso de a referência geográfica a pesquisar se tratar de uma área mais abrangente, como uma cidade ou um país, ignoramos muitas vezes a designação desse local, utilizando-se, somente, o nome que denomina a área. Neste caso, por exemplo, podemos referir “Lisboa” sem especificar que se trata de uma cidade, ou “Portugal” sem pormenorizar que se trata de um país.

Considerando esta prática, sempre que não se proceda à identificação do tipo de localização, no que diz respeito à referência geográfica, então é porque não estamos a contemplar a hipótese de o tipo de localização ser sinónimo de arruamento; assim, os termos da Geo-Net-PT-SSM que têm esta “feature type” associada, não serão contemplados.

Como exemplo, ao desambiguar a referência geográfica “Rua de Lisboa”, serão considerados os termos que possuem o nome “Lisboa” e a “feature type” “Rua”. Já ao desambiguar a referência “Lisboa” serão apenas considerados os termos que possuem o nome “Lisboa” e a “feature type” que não refira qualquer arruamento.

Outro aspecto que se teve em conta na desambiguação de nomes, e que é referido por Gale (Gale et al., 1992), relaciona-se com a existência de mais de uma referência a um nome ambíguo num documento; nestes casos, é provável que o termo ambíguo se refira a um mesmo local geográfico, o que possibilitará a desambiguação desses nomes com a mesma referência ontológica. Assim, e com base nesta lógica, não serão analisadas as referências geográficas que se encontrem repetidas, sendo cada referência geográfica somente considerada uma única vez.

A desambiguação obedeceu, também, ao número de referências geográficas presentes no conjunto de termos a desambiguar. Sob este prisma de análise, a desambiguação tanto pode obter-se de uma única referência geográfica como é possível ser obtida a partir de um conjunto de referências geográficas

4.1.1 Desambiguar uma referência geográfica

A pesquisa de um único termo acarreta uma dificuldade acrescida; de facto, não havendo outras referências geográficas, e que poderiam facilitar a identificação (correcta) do local, é possível que a pesquisa possa corresponder a qualquer termo na ontologia com o mesmo nome.

Como podemos observar na Figura 4.1, se quisermos desambiguar a Referência Geográfica “Rua de Lisboa”, teremos de decidir qual o termo a devolver de entre um conjunto de termos que contêm o nome pesquisado, a associada “feature type” “Rua” e o mesmo valor de IC (como será o caso dos termos de ID: “189102”, “192196”, “374671”).

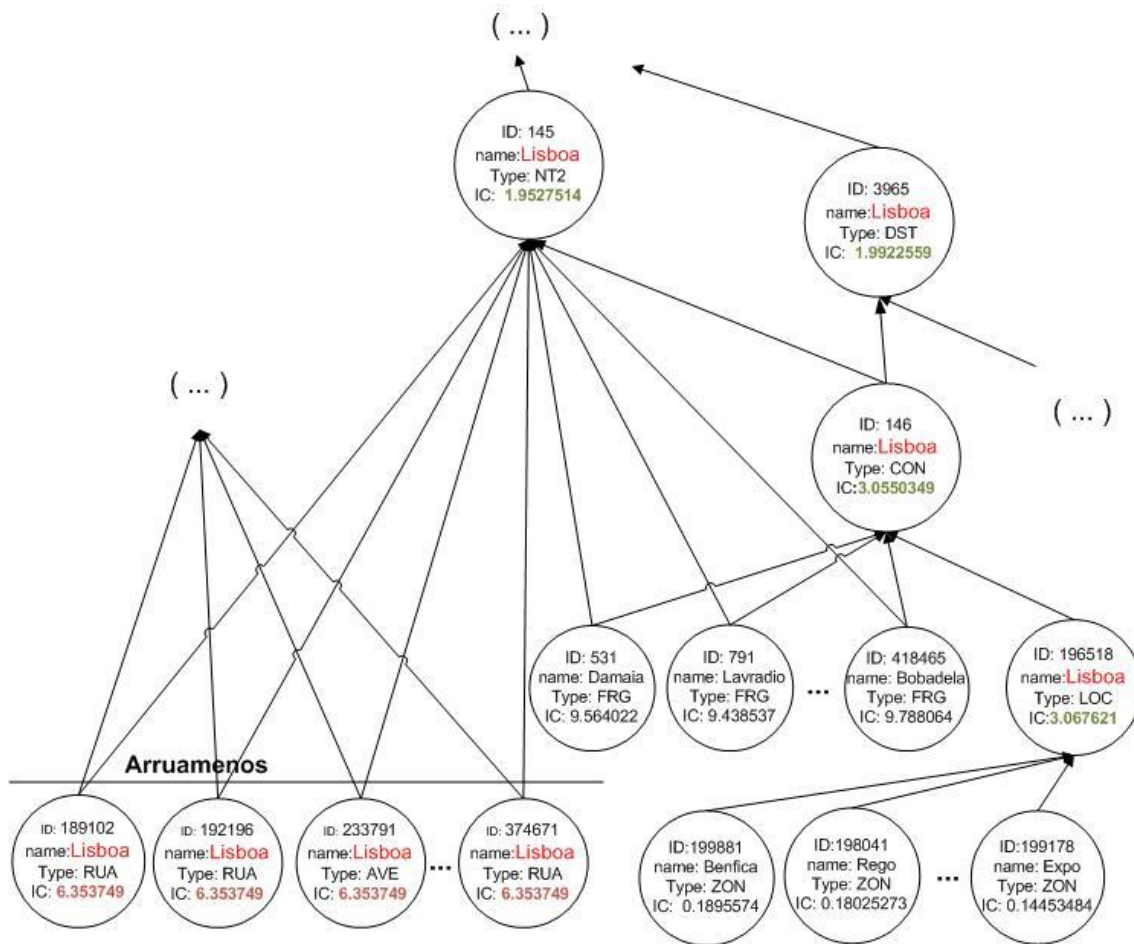


Figura 4.1: Exemplo de diferentes valores de IC para termos de nome ‘Lisboa’ na Geo-Net-PT.

O valor idêntico do IC, nestes termos, justifica-se pelo facto de estarem em causa arruamentos; não haverá, por isso, descendentes a influenciar este valor. Desta forma, o seu IC reflectirá apenas o valor da frequência com que o seu nome associado foi encontrado no “corpus” do Google N-Grams (ver Secção 2.4.2 , relativa ao cálculo do IC).

Já se quisermos desambiguar a referência “Lisboa”, teremos de decidir qual o termo a devolver de entre um conjunto de termos que contêm o nome pesquisado, a “feature type” diferente de qualquer tipo de arruamento e com valor de IC diferenciado, já que estes têm termos sucessores associados, como é o caso dos termos com ID: “145”, “3965”, “146”, ”196518”.

Se o objectivo considerado for a devolução, com maior probabilidade, do termo que corresponde ao que foi efectivamente pesquisado, então as questões referidas nos parágrafos antecedentes assumem especial relevância; será, pois, distinta a forma como abordaremos uma referência geográfica com termos ontológicos com valor de IC distintos e uma referência geográfica com termos ontológicos iguais:

- Termos com valor de IC diferente – Quando se encontram vários termos com o mesmo nome e com um IC associado diferenciado, o termo eleito será aquele que apresenta um IC menor, partindo do princípio que, se um termo possui um IC menor, será porque os seus descendentes são os mais pesquisados na “Internet”, o que torna mais elevada a sua frequência no “corpus”. Por outro lado, ao se devolver o termo com o menor IC, isso significa também que é o mais geral do grupo; se o termo devolvido não corresponder exactamente à pesquisa efectuada, existirá uma maior probabilidade do termo que melhor descreve a referência a desambiguar ser um descendente deste.

- Termos com valor de IC igual – De modo a devolver o termo da ontologia que corresponderá ao efectivamente pesquisado pelo utilizador, o termo que é devolvido não será o que tem o menor IC associado, mas, em alternativa, devolver-se-á o termo com o antecessor de menor IC.

4.1.2 Desambiguação de um conjunto de referências geográficas

A desambiguação é tanto mais útil quanto maior for o número de termos a tratar. Será a partir da análise de todos os termos cujo nome está presente no conjunto de referências a desambiguar, que se tornará possível avaliar quais os termos que têm maior probabilidade de serem os pesquisados.

Procurou analisar-se, sobre este aspecto, a forma mais adequada à desambiguação. Prosseguiram-se várias estratégias, na análise, levando em consideração a diferenciação dos locais. Descrevemos, pois, nesta secção, de forma sucinta, cada estratégia prosseguida e, na secção posterior, elaborar-se-ão demonstrações de cada uma delas.

- **Estratégia 1 (EDNG1)** – Esta é uma estratégia mais complexa a nível da computação, que encontra o conjunto de referências ontológicas que tem a maior média de semelhança semântica de entre todas as combinações possíveis entre termos.
- **Estratégia 2 (EDNG2)** – Esta é uma estratégia que usa heurísticas para desambiguar as referências geográficas. A heurística encontra o âmbito geográfico, referente a dois nomes do conjunto, do par de termos com melhor semelhança semântica entre si. A partir daqui, irá encontrar, para as restantes referências geográficas, o termo mais similar ao âmbito anteriormente escolhido pela estratégia e calcular o âmbito geográfico do novo conjunto de termos.
- **Estratégia 3 (EDNG3)** – Esta estratégia começa por encontrar o âmbito ontológico geográfico comum a todos os termos das referências a desambiguar através de uma heurística. Esta heurística começa por se fixar em duas dessas referências. Encontra, entre os termos de cada uma dessas duas referências, o termo ancestral de nível mais baixo no grafo da ontologia comum a todos esses termos. De seguida, fixa-se noutra referência do conjunto analisado para a obtenção do âmbito geográfico, e identifica o termo ancestral comum ao âmbito encontrado anteriormente, identificando, ainda, todos os termos referentes à nova referência em

análise. Este processo repete-se até que a heurística se tenha fixado em todos os termos ontológicos com nome igual a cada uma das referências geográficas a desambiguar. Uma vez encontrado o âmbito, escolhe, para cada referência geográfica, a referência ontológica que seja mais similar ao âmbito.

- **Estratégia 4 (EDNG4)** – Esta estratégia é semelhante à anterior, pois também começa por definir o âmbito ontológico do conjunto, mas difere na abordagem utilizada para encontrar esse âmbito. Em alternativa, encontra vários conjuntos de termos candidatos a serem o âmbito do conjunto de termos desambiguados, constituídos por, pelo menos, um termo da ontologia representante de cada referência geográfica. Ao proceder desta maneira, eliminam-se, à partida, os termos da Geo-Net-PT-SSM cujo nome de entidade é uma das referências pesquisadas, mas que não iriam obter resultados aceitáveis, no que diz respeito aos outros nomes pesquisados, quanto à semelhança semântica entre os termos. O âmbito escolhido será, então, o do conjunto com a melhor média de similaridade semântica entre os termos desse conjunto. Uma vez encontrado o âmbito, escolhe, para cada referência geográfica, a referência ontológica que seja mais similar ao âmbito.

- **Estratégia 5 (EDNG5)** – Difere da estratégia 4 apenas na forma como selecciona os termos para cada âmbito geográfico ontológico encontrado. Assim, em vez de encontrar os termos da Geo-Net-PT mais semelhantes a esses âmbitos, utiliza a EDNG1 de forma a encontrar o conjunto de termos que tem a maior média de semelhança semântica entre todas as combinações possíveis entre os termos que sejam sucessores de cada um desses âmbitos geográficos ontológicos.

4.1.3 Desambiguação de nomes da Geo-Net-PT: exemplo

Fornece-se, neste ponto, um exemplo que ilustra, de forma concreta, o funcionamento de cada uma das estratégias desenvolvidas neste capítulo.

Este exercício consiste na utilização das estratégias desenvolvidas e de referências geográficas de locais de Portugal que correspondam a mais de um termo na ontologia, como podemos observar na Figura 4.1 As referências geográficas usadas para a estratégia 1, 2 e 3 foram: “Vila Franca de Xira”, “Grande Lisboa” e “Praça Afonso de Albuquerque”. Para as restantes estratégias foram utilizadas as referências: “Lisboa”, “Vila Franca de Xira” e “Praça Afonso de Albuquerque”.

Os termos utilizados em cada demonstração foram escolhidos tendo em conta a sua posição na ontologia por forma a melhor demonstrar o processo de desambiguação de cada uma das referências geográficas em cada estratégia.

<i>Referência Geográfica</i>	<i>IDs Assoc.</i>	<i>IC</i>
Vila Franca de Xira	3552	0.232375
	249101	0.117187
	325	0.095396
Grande Lisboa	129	0.056472
Praça Afonso de Albuquerque	249181	0.238475
	200123	0.238475
Lisboa	379800	6.353749
	196518	3.067621
	146	3.055035
	3965	1.992259
	145	1.952751

Tabela 4.1: Nomes, respectivos IDs e ICs da Geo-Net-PT utilizados na exemplificação das estratégias de desambiguação descritas no Capítulo 4 (GeoScope)

Estratégia EDNG1

Através desta estratégia, obteve-se, em primeiro lugar, todos os termos referentes a cada referência geográfica escolhida (como já mencionado “Vila Franca de Xira”, “Grande Lisboa” e “Praça Afonso de Albuquerque”). Em seguida, constituíram-se subconjuntos de termos relativos a todas as combinações possíveis entre todos os termos de cada uma dessas referências.

Com estes subconjuntos definidos, foi atribuída uma média das semelhanças semânticas entre os seus termos, como é demonstrado na Tabela 4.2.

O cálculo da média de semelhança semântica entre todos estes conjuntos, pode originar, em termos computacionais, uma operacionalização dificultada da estratégia. Não se efectuando uma pré-selecção dos termos utilizados, o sistema irá executar um elevado número de cálculos da semelhança semântica entre termos.

<i>Rank</i>	<i>Scope[Conj. Termos]</i>	<i>Score</i>
1	129 [325, 129, 249181]	0.566026286
2	129 [249101, 129, 249181]	0.564094664
3	129 [325, 129, 200123]	0.488302926
4	129 [249101, 129, 200123]	0.450288336
5	129 [3552, 129, 249181]	0.393050712
6	129 [3552, 129, 200123]	0.33793848

Tabela 4.2: Vários conjuntos de termos da Geo-Net-PT referente a todas combinações possíveis entre os termos da Geo-Net-PT produzidos pela estratégia EDNG1 de desambiguação de nomes.

Como se pode verificar pela Tabela 4.2, o subconjunto com a média mais alta de semelhança semântica, é o usado para desambiguar os nomes do conjunto de referências geográficas ambíguas, sendo atribuído a cada uma dessas referências o termo do subconjunto que o refere na ontologia. Como se pode verificar, o subconjunto com a melhor média de semelhança semântica é o “129 [325, 129, 249181]”.

Estratégia EDNG2

Esta estratégia começou por se fixar nas referências “Grande Lisboa” e “Vila Franca de Xira” e calculou, em primeiro lugar, a semelhança semântica de todas as combinações possíveis entre os termos dessas referências e, de seguida, o âmbito ontológico de cada par. Como se observa na Tabela 4.3, o melhor par de termos para desambiguar estas referências foi o par [325,129], e o termo que melhor descreve o âmbito deste par é o próprio [129] (Grande Lisboa).

<i>Rank</i>	<i>Scope[Conj. Termos]</i>	<i>Score</i>
1	129 [325, 129]	0.743695743
2	129 [249101, 129]	0.650377745
3	129 [3552, 129]	0.391015515

Tabela 4.3: Cálculo da semelhança semântica entre as combinações possíveis de termos referentes às duas primeiras referências geográficas fixadas pela heurística usada em EDNG2.

Fixado este par de termos e o âmbito geográfico ontológico, calculou-se a semelhança semântica entre este âmbito e os termos com nome igual à referência geográfica que se segue no conjunto (“Praça Afonso de Albuquerque”) e o escolhido para desambiguar essa referência, no fundo, o termo mais similar ao âmbito. Como se pode observar na

Tabela 4.4 o termo eleito foi o “249181” e o âmbito geográfico ontológico que descreve o novo conjunto continua a ser o “129” (Grande Lisboa).

<i>Rank</i>	<i>Scope[Conj. Termos]</i>	<i>Score</i>
1	129 [249181, 129]	0.382928378
2	129 [200123, 129]	0.382928378

Tabela 4.4: Cálculo da semelhança semântica entre os termos referentes a “Praça Afonso de Albuquerque” e o âmbito encontrado na iteração anterior.

Assim, o termo que desambigua “Grande Lisboa” é o “129”, o termo que desambigua “Vila Franca de Xira” é o “325” e “Praça Afonso de Albuquerque” é o “249181”, sendo o âmbito geográfico destas referências o “129” (Grande Lisboa).

Estratégia EDNG3

Ao contrário das estratégias previamente utilizadas, aqui define-se, de início, o âmbito geográfico ontológico dos termos referentes a todas as referências geográficas usadas nesta demonstração (“Grande Lisboa”, “Vila Franca de Xira” e “Praça Afonso de Albuquerque”) através de heurísticas.

No cálculo deste exercício, o âmbito encontrado foi o termo “129” (Grande Lisboa) (Tabela 4.5), como sendo o termo mais informativo comum a todos os termos.

Scope escolhido: 129

Tabela 4.5: Âmbito geográfico ontológico escolhido pela heurística usada na Estratégia de Desambiguação de Nomes EDNG3.

Uma vez calculado o âmbito geográfico ontológico para o conjunto das referências geográficas ambíguas, foi escolhido o termo de cada referência que é mais similar a esse âmbito, ou seja, o que apresentou um maior valor de semelhança semântica com este.

Como se pode verificar na Tabela 4.6, os termos escolhidos por esta estratégia para desambiguar os nomes foram uma vez mais o “325”, “129” e “249181”.

<i>Referência Geográfica</i>	<i>ID</i>	<i>SSM(129, ID)</i>
Vila Franca de Xira	3552	0.391015515
	249101	0.650377745
	325	0.743695743
Grande Lisboa	129	1
Praça Afonso de Albuquerque	249181	0.382928378
	200123	0.382928378

Tabela 4.6: Cálculo da semelhança semântica entre os termos das referências geográficas e o âmbito geográfico calculado pela heurística usada na Estratégia de Desambiguação de Nomes EDNG3 (129).

Estratégia EDNG4

A estratégia EDNG4 começou por encontrar todos os termos da Geo-Net-PT-SSM que contivessem, pelo menos, um termo que refira cada nome do conjunto de referências geográficas por desambiguar (termos candidatos a âmbito ontológico dos termos desambiguados). Nesta demonstração foram utilizadas as referências “Lisboa”, “Vila Franca de Xira” e “Praça Afonso de Albuquerque”

Conforme se pode observar na Tabela 4.7, os âmbitos que satisfazem o requisito são os termos “129” (Grande Lisboa) e “3965” (Lisboa) da Geo-Net-PT-SSM.

<i>Âmbitos candidatos</i>	<i>Nome</i>
129	Grande Lisboa
3965	Lisboa

Tabela 4.7: Conjunto de termos candidatos a âmbito geográfico ontológico do conjunto dos termos desambiguados, devolvido pela EDNG4 para os nomes “Vila Franca de Xira”, “Lisboa” e “Praça Afonso Albuquerque”.

Para cada um desses âmbitos e para cada referência geográfica ambígua, foi escolhido o termo que é o mais similar aos âmbitos geográficos ontológicos encontrados. Na Tabela 4.8 e na Tabela 4.9, verificamos que os termos mais similares a “129” (Grande Lisboa) são os termos: “325”, “146” e “249181”, enquanto para o âmbito “3965” (Lisboa) os termos mais similares são: “325”, “3965” e “249181”.

<i>Referência Geográfica</i>	<i>ID</i>	<i>SSM(129,ID)</i>
Vila Franca de Xira	3552	0.391015515
	249101	0.650377745
	325	0.743695743
Lisboa	379800	0.233427849
	196518	0.813262832
	146	0.81524753
	3965	0.482078506
	145	0.963139564
Praça Afonso de Albuquerque	249181	0.382928378
	200123	0.382928378

Tabela 4.8: Cálculo da semelhança semântica entre os termos das referências geográficas e o âmbito ontológico 129 da Geo-Net-PT-SSM.

<i>Referência Geográfica</i>	<i>ID</i>	<i>SSM(3965,ID)</i>
Vila Franca de Xira	3552	0.374390921
	249101	0.627023302
	325	0.718775041
Lisboa	379800	0.236503387
	196518	0.787472106
	146	0.789435733
	3965	1
	145	0.500343442
Praça Afonso de Albuquerque	249181	0.366569324
	200123	0.366569324

Tabela 4.9: Cálculo da semelhança semântica entre os termos das referências geográficas e o âmbito ontológico 3965 da Geo-Net-PT-SSM.

Uma vez encontrados os termos mais similares para cada âmbito, calculou-se a média de semelhança semântica entre esses termos. Escolheu-se, de seguida, para desambiguar as referências geográficas o conjunto que apresentou a média de semelhança semântica mais alta entre os termos do conjunto.

Nesta estratégia, o conjunto escolhido foi “ 39659 [325, 3965, 249181] ” para a desambiguação dos nomes (ver

Tabela 4.10).

<i>Rank</i>	<i>Scope[Conj. Termos]</i>	<i>Score</i>
1	3965 [325, 3965, 249181]	0.552266368
2	129 [325, 145, 249181]	0.547214991

Tabela 4.10: Médias de semelhança semântica entre os termos dos conjuntos produzidos pela EDNG4.

Estratégia EDNG5

Como anteriormente acontecera com a EDNG4, a estratégia EDNG5 iniciou-se com a pesquisa dos vários termos da Geo-Net-PT-SSM que tivessem, pelo menos, um termo sucessor que referenciasse cada nome ambíguo (Tabela 4.7).

A estratégia EDNG4, relembre-se, procurava os termos de cada referência geográfica mais similares a cada âmbito; por oposição, a EDNC5 elegeu os termos que eram mais similares entre si em cada conjunto encontrado (considerando o âmbito e os termos).

Como se pode verificar na Tabela 4.11, o âmbito geográfico ontológico 129 tem os termos [325, 3552, 249101] que referenciam “Vila Franca de Xira”, [146, 196518] que referenciam “Lisboa” e [200123, 249181] que referenciam “Praça Afonso de Albuquerque”. Já para o âmbito geográfico ontológico 3965 foram encontrados os conjuntos de termos [325, 3552, 249101], [3965] e [200123, 249181] respectivamente.

<i>Referências</i>	<i>Ancestor(129)</i>	<i>Ancestor (3965)</i>
Vila Franca de Xira	[325, 3552, 249101]	[325, 3552, 249101]
Lisboa	[146, 196518]	[3965]
Praça Afonso de Albuquerque	[200123, 249181]	[200123, 249181]

Tabela 4.11: Termos sucessores de cada termo candidato a âmbito geográfico ontológico do conjunto dos termos desambiguados.

Após a pesquisa de termos que referenciam cada nome nos respectivos âmbitos, utilizou-se a primeira estratégia desta secção de modo a encontrar a combinação de termos mais similares. Como se pode observar na Tabela 4.12 e na Tabela 4.13. as melhores combinações são a 129 [249101, 146, 249181] e a 3965[325, 3965, 249181] respectivamente.

<i>Rank</i>	<i>Scope[Conj. Termos]</i>	<i>Score</i>
1	129 [249101, 146, 249181]	0.526054
2	129 [249101, 196518, 249181]	0.52561
3	129 [325, 146, 249181]	0.52008
4	129 [325, 196518, 249181]	0.519553
5	129 [325, 196518, 200123]	0.495709
6	129 [325, 146, 200123]	0.49559
7	129 [249101, 196518, 200123]	0.465684
8	129 [249101, 146, 200123]	0.465481
9	129 [3552, 146, 249181]	0.372249
10	129 [3552, 196518, 249181]	0.371997
11	129 [3552, 196518, 200123]	0.370764
12	129 [3552, 146, 200123]	0.37037

Tabela 4.12: Combinações geradas pela estratégia EDNG1 para os sucessores de 129 e respectiva média de semelhança semântica entre os seus termos.

<i>Rank</i>	<i>Scope[Conj. Termos]</i>	<i>Score</i>
1	3965 [325, 3965, 249181]	0.552266
2	3965 [249101, 3965, 249181]	0.550857
3	3965 [325, 3965, 200123]	0.474543
4	3965 [249101, 3965, 200123]	0.437051
5	3965 [3552, 3965, 249181]	0.382056
6	3965 [3552, 3965, 200123]	0.326944

Tabela 4.13: Combinações geradas pela estratégia EDNG1 para os sucessores de 3965 e respectiva média de semelhança semântica entre os seus termos.

Para efectuar a desambiguação das referências geográficas, escolheram-se, enquanto referências ontológicas, os termos contidos no conjunto com melhor semelhança semântica entre os restantes termos. Neste caso, os eleitos foram o “3552”, “146” e “200123” (Tabela 4.14)

<i>Rank</i>	<i>Scope[Conj. Termos]</i>	<i>Score</i>
1	129 [3552, 146, 200123]	0.37037
2	3965 [3552, 3965, 200123]	0.326944

Tabela 4.14: Médias de semelhança semântica entre os termos dos conjuntos produzidos pela estratégia EDNG5.

Uma vez que conseguimos realizar a desambiguação de um conjunto de referências geográficas através da semelhança semântica, podemos afirmar que estão lançadas as bases metodológicas para a desambiguação de referências geográficas extraídas de documentos e a respectiva atribuição de um âmbito geográfico a estes, factor essencial para dotar os Recuperador de Informação Geográfico de conhecimento ontológico.

Então, o processo de decisão quanto às páginas que serão apresentadas numa pesquisa envolverá o cálculo da similaridade entre cada página e a respectiva pesquisa, recorrendo, para isso, ao uso das medidas de semelhança semântica. Serão, por isso, implementadas várias estratégias diferentes que produzem um valor de similaridade entre conjuntos de referências geográficas existentes nas páginas e a partir das referências ontológicas referentes a cada uma delas.

Contudo, e pretendendo-se que esta pesquisa devolva resultados em concordância com a ontologia geográfica, permanecem diversos problemas que necessitam de uma

solução e, ao mesmo tempo, persistem algumas constatações que devem ser observadas. Em primeiro lugar, o sistema (GIR) terá que fazer a ligação entre as referências geográficas, e que estão a ser testadas no processo de avaliação de similaridade, e os termos contidos na ontologia geográfica que as representam. Esta ligação é conseguida, por um lado, através da atribuição de um âmbito geográfico ontológico a cada conjunto de termos geográficos; por outro, obtém-se pela desambiguação de cada uma das referências geográficas. A necessidade de desambiguação explicar-se-á pela existência de vários locais em Portugal com o mesmo nome, o que conduz à impossibilidade de definir directamente a que locais se refere cada termo da pesquisa.

4.2 Caso de Estudo

Nesta secção são apresentados exemplos de como as medidas de desambiguação implementadas ao longo deste projecto podem ser utilizadas para desambiguar resumos ricos em referências geográficas. Para a realização destes exemplos foram utilizadas páginas, que se encontram organizadas por regiões, do “site” “Portugal Tribe”⁸: “Porto e Norte De Portugal”, “Beiras”, “Lisboa e Vale Do Tejo”, “Alentejo”, “Algarve”, “Açores” e “Madeira”. Podem ser observados na Tabela 4.15. os termos que foram seleccionados por página; a selecção manual obedeceu a um padrão de triagem que tem em conta o texto que a compõe, nomeadamente as referências geográficas ali contidas.

<i>Âmbito Geográfico da pag. Web</i>	<i>Referências Geográficas extraídas manualmente</i>
PortoeNorteDePortugal	[Porto, Norte, Torre de Moncorvo, Caminha, Amarante, Peso da Régua]
Beiras	[Coimbra, Figueira da Foz, Vagos, Penedono, Seia]
LisboaeValeDoTejo	[Lisboa, Alcochete, Sesimbra, Arruda dos Vinhos, Montijo]
Alentejo	[Alentejo, Alter do Chão, Moura, Monsaraz, Évora, Barrancos]
Algarve	[Algarve, São Brás de Alportel, Olhão, Monchique, Vila Real de Santo António, Faro, Santa Barbara de Nexe]
Açores	[Ilha da Graciosa, Santa Cruz da Graciosa, Nossa Senhora da Ajuda, São Salvador, São João, Ilha Terceira]
Madeira	[Madeira, Ilha de Porto Santo, Porto Santo, Ponta da Calheta]

Tabela 4.15: Referências geográficas extraídas manualmente das páginas Web do site Portugal Tribe.

⁸ <http://mytribe.portugaltribe.com/>

A partir das anotações geográficas retiradas, utilizou-se o GeoSSM para o cálculo, através das medidas de semelhança semântica, do âmbito geográfico de cada página, encontrando-se, assim, os termos da ontologia utilizada que correspondem a cada uma das anotações extraídas. Por fim, os âmbitos geográficos calculados pelo sistema implementado foram comparados ao âmbito geográfico presente em cada página (ver Tabela 4.16).

Através destes exemplos foi possível constatar que todas as medidas de desambiguação procederam à desambiguação de termos de um determinado conjunto de termos de âmbito conhecido.

<i>Site</i>	<i>AGO</i>	<i>[Ref.Geográfica:RO]</i>
www.PortoeNorteDePortugal.pt	196	[porto:3967], [norte:196], [torre de moncorvo:301], [caminha:71], [amarante:32], [peso da regua:224]
www.Beiras.pt	94	[coimbra:3949], [figueira da foz:117], [vagos:308], [penedono:220], [seia:273]
www.LisboaeValeDoTejo.pt	418732	[lisboa:146], [alcochete:8], [sesimbra:279], [arruda dos vinhos:43], [montijo:182]
www.Alentejo.pt	12	[alentejo:12], [alter do chao:26], [moura:186], [monsaraz:2740], [evora:3950], [barrancos:53]
www.Algarve.pt	17	[Algarve:17], [sao bras de alportel:289], [olhao:202], [monchique:176], [vila real de santo antonio:336], [faro:113], [santa barbara de nexe:1351]
www.Acores.pt	418745	[ilha da graciosa:3952], [santa cruz da graciosa:141847], [nossa senhora da ajuda:185837], [sao salvador:6787], [sao joao:418462], [ilha terceira:3962]
www.Madeira.pt	418745	[Madeira:241], [ilha de porto santo:3955], [porto santo:156680], [ponta da calheta:-1]

Tabela 4.16: Esta tabela representa as RO (referências ontológicas) e respectivos AGO (âmbito geográfico ontológico) calculados pelas diferentes medidas de semelhança semântica com as diferentes EDNG.

No entanto, podemos observar na tabela inferior (Tabela 4.17) que, salvo algumas exceções, o antecessor comum a todos os termos gerados pelas medidas de semelhança semântica, o qual define o âmbito do conjunto, obteve algum grau de adequação ao âmbito das páginas; o que nos leva a defender que estes resultados são satisfatórios no processo de desambiguação de referências geográficas através das medidas implementadas.

<i>Pág Portugal Tribe</i>	<i>ID</i>	<i>T_id</i>	<i>Nome</i>
Porto e Norte de Portugal	196	NT2	Norte
Beiras	94	NT3	Continente
Lisboa e Vale do Tejo	418732	PRO	Estremadura
Alentejo	12	NT2	Alentejo
Algarve	17	NT3	Algarve
Açores	252	NT3	Região Autónoma dos Açores
Madeira	418745	PAI	Portugal

Tabela 4.17: Âmbitos Geográficos Ontológicos das páginas do site Portugal Tribe e respectivos âmbitos geográficos ontológicos calculados.

A título de exemplo, e no caso da Madeira, o âmbito geográfico ontológico encontrado foi a raiz da ontologia (ID: 418745, t_id: PAI, n_name: Portugal). Tal ficou a dever-se à não existência, na ontologia utilizada, de um termo com um nome idêntico a uma das referências geográficas contidas no conjunto de termos associados à ilha da Madeira (Calheta). Na realidade, detectamos a existência de um local, na Madeira, com esse nome, o Concelho da Calheta; no entanto, o termo que o define na Geo-Net-PT-SSM utiliza um nome alterado, incluindo “Madeira” junto ao seu nome: “Calheta (Madeira)”. Isto possibilita, pois, que a desambiguação, efectuada pelo sistema desenvolvido, possa não ser correctamente efectuada, ou mesmo impossível, com alguns nomes de locais; de facto, isso pode acontecer na desambiguação de referências geográficas presentes nos documentos quando, na Geo-Net-PT-SSM, não haja um termo correspondente na ontologia, ou o nome dessa referência não seja o usado pelo termo que descreve essa referência na ontologia.

4.2.1 Utilização do trabalho desenvolvido

No seguimento do trabalho desenvolvido durante este projecto, e ainda no âmbito do projecto GREASE, Batista (S. Batista, David et al., 2010), desenvolveu um sistema capaz desambiguar de uma forma automática, referências geográficas presente em documentos de texto. Esta desambiguação foi realizada através GeoSSM e GeoScope desenvolvidos no presente projecto.

Nesse projecto, a desambiguação de termos através das medidas de semelhança semântica foram testados em larga escala, de modo a ser provado a sua relevância para servir este propósito.

Para a realização dos testes referidos, foram retiradas automaticamente referências geográficas presentes nos textos da Wikipédia Portuguesa (www.wikipedia.pt) através do Minotird. De seguida, realizou-se a desambiguação dessas referências extraídas através dos métodos de desambiguação por mim desenvolvidos. Por fim, e entre toda a amostra retirada, avaliaram-se quantas referências foram correctamente desambiguadas.

Tal como foi apontado neste projecto, também David (S. Batista, David et al., 2010) encontrou, no processo de desambiguação de termos, problemas que foram referidos anteriormente no presente projecto e que se prendem com o facto de existirem referências encontradas nos documentos analisados, que se querem desambiguar, que não constam na Geo-Net-Pt.

Page of	Correctly Extracted	Correctly Disambiguated
Aveiro	100%	70%
Beja	88%	87%
Braga	71%	67%
Bragança	100%	75%
Castelo Branco	38%	54%
Coimbra	70%	82%
Évora	100%	100%
Faro	80%	68%
Guarda	93%	76%
Leiria	90%	85%
Lisboa	96%	92%
Portalegre	90%	68%
Porto	87%	68%
Santarém	100%	81%
Setúbal	81%	70%
Viana do Castelo	100%	62%
Vila Real	77%	83%
Viseu	92%	89%

Tabela 4.18: Percentagem de entidades correctamente extraídas e percentagem das entidades correctamente desambiguadas.

Fonte: Batista et al. 2010

Ainda no âmbito do projecto GREASE foi criada uma API, a Geographical Similarity Calculator GeoSSM⁹ que teve por base a GeoSSM. A GeoSSM implementa as medidas de semelhança semântica aplicadas a uma ontologia geográfica, bem como a ontologia Geo-Net-PT-SSM. Tal como a GeoSSM, também a Geo-Net-PT-SSM foi criada durante o desenvolvimento deste projecto.

⁹ Esta API encontra-se disponível no “site”
http://xldb.fc.ul.pt/wiki/Geographic_Similarity_calculator_GeoSSM

Capítulo 5 Conclusões

O presente trabalho foi desenvolvido no âmbito do projecto GREASE e estudou a implementação de medidas de semelhança semântica entre termos de uma ontologia geográfica baseada na segunda versão da Geo-Net-PT; esta ontologia descreve as divisões geográficas do território português e a pesquisa desenvolvida teve como principal objectivo/pressuposto o estudo das vantagens/utilidade que as medidas de semelhança semântica podem trazer às pesquisas geográficas na Web. Pretendia-se alcançar dois objectivos distintos:

- Implementação de medidas de semelhança semântica no âmbito de uma ontologia geográfica;
- Desambiguação de um conjunto de referências geográficas através das medidas de semelhança implementadas e posterior atribuição do respectivo âmbito geográfico a esses conjuntos.

Implementaram-se, assim, duas ferramentas, o GeoSSM e a GeoScope. O GeoSSM, convirá salientar, constituiu a implementação principal deste projecto; implementou, assim, as medidas de semelhança semântica no âmbito de uma ontologia geográfica. Por sua vez coube ao GeoScope a capacidade de retirar ou de minimizar a ambiguidade de um conjunto de referências geográficas (Âmbito Geográfico Ontológico e Referências Ontológicas) relacionadas com esse sumário.

5.1 Experiências com as várias Medidas de Semelhança Semântica

Forneceram-se, com este estudo, vários exemplos das medidas de semelhança semântica implementadas, o que permitiu apurar algumas vantagens e desvantagens associadas a cada uma destas medidas. Em acréscimo, no cálculo da semelhança semântica considerou-se o número de antecessores comuns observados por cada par, o MICA respectivo, o valor do GRASM e os subgrafos produzidos pela união e intersecção dos antecessores de cada termo do par.

Foi possível constatar, a partir dos exemplos de verificação das medidas de semelhança semântica, as mesmas limitações, já apontadas em estudos anteriores, aplicadas à ontologia utilizada no presente projecto. De facto, a explicação para essas limitações de análise estará relacionada com a elevada percentagem de termos com o mesmo valor de IC contidos na ontologia e, adicionalmente, com a ausência de representatividade, nessa ontologia, de muitas organizações diferentes do território de Portugal.

Não obstante, propôs-se, nestas páginas, uma nova abordagem para o cálculo da semelhança semântica entre termos; acreditamos que poderá vir a fornecer uma resposta válida aos problemas encontrados nas medidas de semelhança semântica implementadas na ontologia geográfica. Esta nova abordagem, apresentada na secção 3.4, traduziu-se na média entre a soma das duas medidas conhecidas, a Sim_{Gic} e a $Sim_{JiangConrath}$.

5.2 Desambiguação do Âmbito Geográfico

Foram implementados diferentes algoritmos de desambiguação dos termos da Geo-Net-PT-SSM. No entanto, a impossibilidade de utilização do processo de teste previsto provocou a não realização de um estudo aprofundado sobre estes algoritmos.

Como foi possível comprovar, todavia, as medidas de semelhança semântica podem servir para fazer a desambiguação de termos com nomes idênticos, principalmente se na ontologia usada não forem predominantes os termos com IC de valor igual; este facto, aliás, revelou-se prejudicial a algumas das medidas de semelhança semântica implementadas.

Apesar de não se ter conseguido encontrar a melhor métrica para produzir semelhança semântica entre os termos da ontologia Geo-Net-PT-SSM, os testes apresentados dão boas indicações de que as medidas de semelhança semântica possam ser integradas, de futuro, num sistema de recuperação de informação geográfica dotado de conhecimento ontológico.

5.3 Geo-Net-PT

Ao longo deste projecto foram identificados alguns dos problemas relacionados com a ontologia utilizada para o cálculo de semelhança semântica, nomeadamente no que diz respeito ao nível de estrutura e dos nomes utilizados nos vários termos que a compõem.

Uma das limitações encontradas ficou a dever-se à existência de problemas nos nomes utilizados em alguns dos termos, como é o caso do termo (id: 69) que se refere ao concelho de Calheta no Arquipélago da Madeira e que apresenta um nome na Geo-Net-PT 02 com informação adicional - i.e., “Calheta (Madeira)”. O mesmo aconteceu para um local com o mesmo nome no Arquipélago dos Açores – “Calheta (Açores)”. Este tipo de alteração ao nome da entidade geográfica impossibilita que se lhe associe uma referência directa e simples.

Outra limitação da ontologia usada para o cálculo de semelhança semântica entre os seus termos, diz respeito à forma como a Geo-Net-PT referencia os nomes alternativos das entidades geográficas. Aqui são adicionadas novas características à ontologia e é estabelecida uma relação de “part-of” entre o termo alternativo e o nome comum da mesma referência geográfica. Desta maneira, as medidas de semelhança semântica vão assumir que estas traduzem características diferentes, produzindo um valor de semelhança semântica diferente entre uma referência ontológica e si mesmo

Detectou-se, por fim, que as freguesias não têm uma relação de antecessor com os arruamentos (que são seus sucessores no mundo real), como se pode observar na figura referente às relações do tipo “Part of” dos termos da Geo-Net-PT (Figura 2.3).

5.4 Trabalho Futuro

O desenvolvimento deste projecto trouxe, naturalmente, um conjunto de ideias que poderia justificar o aprofundamento da análise e o acréscimo de trabalho subsequente.

Em primeiro lugar, justificar-se-ia a utilização de um novo “corpus” para o cálculo do IC. De facto, e como foi referido ao longo deste relatório, denotou-se a existência de um grande número de termos na ontologia que partilham o mesmo valor de IC. Tal deve-se ao facto ter sido utilizado um “corpus” que expressa a ocorrência de palavras da língua inglesa e que, por essa razão, não contém muitos dos nomes de locais portugueses. A escolha do “corpus” utilizado no cálculo do IC incidiu sobre o Google-Ngrams; não existe, de momento, uma ferramenta semelhante, e exclusiva, para a Web portuguesa ou, tão-só, para as palavras de língua portuguesa. Irá no entanto ser desenvolvido pelo GREASE um “corpus” com essas características que irá utilizar a colecção de documentos da Web portuguesa, a WPT 05¹⁰. Daqui se depreenderá que uma proposta de trabalho futuro passaria pelo cálculo do novo IC de cada termo a partir de um novo “corpus”, produzindo-se melhores valores de semelhança semântica entre termos.

Tivemos, ainda, a oportunidade de referir, neste relatório, os problemas encontrados com a ontologia utilizada. Assim, e de futuro, seria útil melhorar a ontologia geográfica, criando-se uma nova versão da Geo-Net-PT-SSM a partir de novas versões da Geo-Net-PT, considerando, e licitamente, que esta ontologia se encontra em constante desenvolvimento e processo de melhoria.

Por outro lado, em termos de trabalho a desenvolver no futuro, seria relevante melhorar a adaptação da ontologia geográfica às SSM; descrevemos, a propósito, neste relatório, alguns dos problemas que encontrámos no que diz respeito à implementação das medidas de semelhança semântica. Justificar-se-ia a criação de uma nova versão da Geo-Net-PT-SSM, desenvolvida a partir de novas versões da Geo-Net-PT, considerando que se trata de uma ontologia em constante aperfeiçoamento. Assim, poder-se-ia dotar a ontologia de organizações alternativas do território português para que seja possível a melhor aplicabilidade de medidas que usam termos disjuntivos para calcular a semelhança semântica entre termos.

A implementação de novos algoritmos constituiria outro eixo prioritário de trabalho futuro. De facto, tornar-se-ia essencial implementar e testar novos algoritmos de desambiguação de termos e algoritmos de semelhança entre contextos geográficos que

¹⁰ http://xldb.fc.ul.pt/wiki/WPT_05

tenha em conta “outliers” pois temos que ter em conta que as páginas “Web” podem fazer uma referência esporádica a um termo fora do âmbito geográfico desta.

Por último, o trabalho futuro poderá ainda enveredar pela realização de novos testes que comprovem qual a medida de semelhança semântica que melhor se adequa à ontologia geográfica Geo-Net-PT.

Bibliografia

Bruno Martins. *Geographically Aware Web Text Mining*. Tese de Doutoramento, Universidade de Lisboa, Faculdade de Ciências, Departamento de Informática, Agosto 2008.

Bruno Martins. Geographical Information Retrieval. In *Proc.of the Seminário Língua Natural, Lisboa, Tagus Park, 2009*

Cátia Pesquita. *Improving Semantic Similarity for Proteins based on the Gene Ontology*. Tese de Mestrado, Faculdade de Ciências, Departamento de Informática, Universidade de Lisboa, 2006

Catia Pesquita, Daniel Faria, André Falcão, Philip Lord, Francisco Couto. Semantic Similarity in Biomedical Ontologies. *PLoS Computational Biology*. 2009

Claudia Leacock, and Martin Chodorow. Combining local context with wordnet similarity for word sense identification. In *WordNet: A Lexical Reference System and its Application*, 1998.

Daniel Faria, Catia Pesquita, Francisco Couto, André Falcão, ProteInOn: A Web Tool for Protein Semantic Similarity. Relatório Técnico. TR 07-6. DI FCUL, March 2007

David Batista. *Prospecção de Conceitos Geográficos na WEB*. Tese de Mestrado, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, 2009

David Batista, Mário J. Silva, Francisco Couto and Bibek Behera. Geographic Signatures for Semantic Retrieval. In *GIR'10: In Proc of the 6th Workshop on Geographic Information Retrieval Zurich, Switzerland, 18-19 Fevereiro 2010*

Dekang Lin. An information-theoretic definition of similarity. In *Proc. Of the 15th International Conference on Machine Learning*. São Francisco, CA, Estados Unidos da América. Julho de 1998.

Fonseca, F., Sheth, A.: *The Geospatial Semantic Web*. UCGIS White Paper. (2002). Available at <http://www.ucgis4.org/priorities/research/2002researchagenda.htm>

Francisco Couto. *ReBIL: Relating Biological Information through Literature*. Tese de Doutoramento, Faculdade de Ciências, Departamento de Informática, Universidade de Lisboa, 2006

Francisco J. Lopez-Pellicer, Marcirio Chaves, Catarina Rodrigues, Mário J. Silva. Geographic Ontologies Production in Grease-II Relatório Técnico. TR 09-18. University of Lisbon, Faculty of Sciences, LASIGE, November 2009.

Jay J. Jiang, and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the 10th International Conference on Research on Computational Linguistics*. 1997

Karla A.V. Borges, Alberto H. F. Laender, Cláudia B. Medeiros, and Clodoveu A. Davis, Jr. Discovering geographic locations in web pages using urban addresses. In *GIR '07: In Proc of the 4th ACM workshop on Geographical information retrieval, pages 31–36, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-828-2*.

Marcirio Chaves, Mário J. Silva, and Bruno Martins. A geographic knowledge base for semantic web applications. In *20th Brazilian Symposium on Databases - SBDD, pages 40–54, Uberlândia, Minas Gerais, Brazil, October 2005*.

Marcirio Chaves, Catarina Rodrigues, and Mário Silva. Data model for geographic ontologies generation. In *XATA 2007 - XML: Aplicações e Tecnologias Associadas, February 2007*.

Marcirio Silveira Chaves. *Uma Metodologia para Construção de Geo-Ontologias*. Tese de Doutorado, Faculdade de Ciências, Universidade de Lisboa, 2009.

Martin Warin, Henril Oxhammar, and Martin Volk. Enriching an Ontology with WordNet based on Similarity Measures. In *Proc of the Meaning 2005 Workshop*. Trento, Itália, Fevereiro 2005

Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of the 14th International Joint Conference on Artificial Intelligence*. Montreal, Canadá, Agosto 1995

Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. In *IEEE Transaction on Systems, Man, and Cybernetics*, volume 1.

Renata Viegas. *GeOntoQuery – Um Mecanismo de Busca em Bancos de Dados Geográficos Baseado em Ontologias*. Tese de Mestrado, Departamento de Informática e Matemática Aplicada, Universidade Federal do Rio Grande do Norte, Brasil, 2006.

Robert Gentleman. Visualizing and Distances Using GO, 2005, Disponível em: <http://www.bioconductor.org/repository/devel/vignette/GOvis.pdf>

Tom Gruber. The role of common ontology in achieving sharable, reusable knowledge bases. In *Allen, J. F., Fikes, R., and Sandewall, E., editors, KR'91: Principles of Knowledge Representation and Reasoning*, Morgan Kaufmann, San Mateo, California, 1991.

Tom Gruber. “Ontology” in *Encyclopedia of Database Systems*, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2008

Tran Minh Duc, Toshikazu Nishimura. Geographical Information Retrieval System using Semantic Relationships between Multiple Layers, In *iiWAS '08: Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services*, Linz, Austria, 2008.

William A. Gale, Kenneth W. Church, and David Yarowsky. One sense per discourse. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pages 233–237, Morristown, NJ, USA, 1992. Association for Computational Linguistics. ISBN1-55860-272-0. doi: <http://dx.doi.org/10.3115/1075527.1075579>.

Zhibiao Wu, and Martha Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd. Annual Meeting of the Association for Computational Linguistics (ACL 1994)*