

# ***Apoio à negociação conceptual com base em processos híbridos de avaliação de similaridade semântica***

Luís Costa <sup>1</sup>, Carla Pereira <sup>1,2</sup>, Cristóvão Sousa <sup>1,2</sup>

1) CIICESI, ESTGF, Instituto Politécnico do Porto, Portugal  
[8050120@estgf.ipp.pt](mailto:8050120@estgf.ipp.pt)

2) INESC TEC, Porto, Portugal  
[csp@estgf.ipp.pt](mailto:csp@estgf.ipp.pt), [cds@estgf.ipp.pt](mailto:cds@estgf.ipp.pt)

## **Resumo**

A modelação conceptual permite que especialistas descrevam um determinado domínio em estruturas do tipo conceito - relação - conceito. Num ambiente colaborativo, para um mesmo domínio, um processo de conceptualização pode originar várias soluções. É necessário que os participantes cheguem a um consenso que represente uma solução partilhada. O foco principal deste trabalho situa-se nesta etapa. Pretendemos criar uma nova abordagem para integração de estruturas conceptuais de diferentes modelos, facilitando assim o processo de negociação. Para tal, este trabalho recorre à utilização de algoritmos de similaridade semântica.

**Palavras-chave:** Similaridade semântica, modelação conceptual, conceptualização partilhada

## **1. Introdução**

Pereira [Pereira 2010] propôs um método de apoio à construção colaborativa de artefactos semânticos e afirma que a criação destes artefactos atua como um mecanismo sociotécnico, defendendo o ponto de vista de que o significado é construído socialmente através da colaboração e negociação. As redes colaborativas caracterizam-se pela sua multidisciplinaridade e multiculturalidade incorporando visões do mundo diversificadas, sendo natural a existência de problemas de correspondência conceptual. As partes envolvidas devem, por isso, identificar os conceitos e relações e acordar quanto ao uso e representação desses elementos. Este processo é longo e árduo, porque: a) os especialistas têm dificuldade em explicitar o seu conhecimento; b) em grandes repositórios de conhecimento, torna-se difícil acompanhar as modificações propostas, as razões que sustentam essas modificações e as suas repercussões no modelo conceptual final; c) há a necessidade de estabelecer consenso entre várias pessoas e d) a informação e o seu significado são dependentes do tempo e contexto.

Evoluindo o trabalho realizado em [Pereira 2010] e [Pereira *et al.* 2012], pretendemos acrescentar novos mecanismos que melhorem a negociação conceptual [Druckman 2007] durante a criação de modelos conceptuais partilhados.

A tarefa de negociação visa a obtenção de consenso em cenários que denunciem situações de conflito e/ou ambiguidade no uso dos termos para representação de domínios específicos.

Num cenário de tratamento de conflitos e repetições de termos, a similaridade semântica surge como um conceito fundamental. Segundo Lin [Lin 2000], a similaridade semântica pode ser interpretada como o grau de proximidade taxonómica entre termos. As medidas de similaridade retornam um valor que quantifica a proximidade em função dos traços semânticos observados numa ou várias fontes de conhecimento [Sánchez & Batet 2012]. O valor obtido do cálculo de similaridade entre conceitos vai permitir identificar elementos com o mesmo significado, mesmo aqueles que possuem léxico diferente. Com o nosso trabalho, pretendemos desenvolver uma nova abordagem para identificação de similaridade semântica entre modelos acreditando que se trata de uma contribuição efetiva para melhorias de eficiência e eficácia do processo de criação de um modelo conceptual partilhado.

O cálculo da similaridade semântica pode ser efetuado segundo diferentes medidas e abordagens. Na literatura classificam as medidas segundo quatro grandes categorias [Petrakis *et al.* 2006]: 1) medidas baseadas em ontologias; 2) medidas baseadas em *Information Content (IC)*; 3) medidas baseadas em características e 4) medidas híbridas. As medidas baseadas em ontologias, ou também chamadas de “*path length*”, têm por princípio comum o cálculo da similaridade através do tamanho do caminho que liga dois conceitos de acordo com a sua posição numa taxonomia. Nas medidas baseadas em *IC* a similaridade semântica é obtida com base no teor de informação de cada conceito. Quanto mais informação comum dois conceitos partilharem mais similares eles são. O *WordNet*<sup>1</sup> surge como fonte de informação principal, contudo, não disponível para todas as línguas. As medidas baseadas em características são independentes da taxonomia e baseiam-se na suposição que cada conceito é descrito por um conjunto de palavras que indicam as suas propriedades ou características, tais como as suas definições ou “*glosses*” no *WordNet*. Quanto mais características em comum dois conceitos possuírem, e menos características não comuns, mais similares os conceitos serão. No quarto e último grupo, as medidas híbridas combinam as ideias acima apresentadas. Na prática, tentam incluir no cálculo de similaridade medidas que olhem aos sinónimos (*IC*), vizinhança (taxonomia) e características dos conceitos. A cada tipo de medida é atribuído um valor que no final é somado e obtém-se o valor global de similaridade.

Das diversas áreas de aplicação da similaridade semântica destacam-se como principais: recuperação de informação; integração de dados; integração de base dados; deteção de ambiguidades e duplicação de informação.

Na secção 2 é apresentado a nossa abordagem de utilização das medidas de similaridade semântica ao problema de negociação conceptual. Em seguida, na secção 3, são analisados algumas medidas existentes para perceber as vantagens e desvantagens da sua utilização. Por último, enunciamos as conclusões obtidas até ao momento e os próximos passos.

---

<sup>1</sup> <http://wordnet.princeton.edu>

## 2. Processo de similaridade semântica

O nosso trabalho pretende ser um contributo para facilitar os especialistas na elaboração de modelos conceptuais de forma colaborativa, mais especificamente, na fase de negociação conceptual, no qual podem existir conceitos repetidos ou com o mesmo significado.

A modelação conceptual partilhada possui uma serie de fatores (indicados na secção 1 de acordo com Pereira) que aumentam a complexidade no momento de chegar a um consenso na criação do modelo final. Para superar este problema apresentamos um serviço que visa detetar conceitos similares em modelos diferentes para uma posterior junção num único conceito. Este processo de cálculo de similaridades e “merge” de conceitos permite uma melhoria da eficiência e eficácia no processo de criação de um modelo conceptual final.

A plataforma ConceptME<sup>2</sup> [Costa *et al.* 2012] apresenta-se como uma solução para a criação de modelos conceptuais de forma colaborativa. Durante um processo de conceptualização usando o ConceptME obtemos um conjunto de informação a ter em conta nas análises de similaridade. Como principais recursos disponíveis num modelo conceptual temos: i) estruturas conceito-relação-conceito (C-R-C) em que cada elemento possui as suas propriedades semânticas. Os conceitos além do nome possuem uma lista de variantes e a sua definição. As relações possuem o nome / *label* e o tipo ou categoria de relação; ii) Corpus indexado; e iii) Catálogo relações.

A Figura 1 ilustra o processo criado para o cálculo da similaridade de acordo com o nosso problema e recursos disponíveis.



Figura 1 – Processo de similaridade semântica

<sup>2</sup> <http://www.conceptme.pt>

A nossa abordagem consiste na aplicação de diferentes tipos de métricas para o cálculo de similaridade semântica. Os resultados de similaridade serão obtidos ao fim de três passos. No primeiro serão aplicadas técnicas para transformação e normalização dos nomes usados nos conceitos e relações. No segundo passo é efetuado uma análise sintática para comparação de conceitos e relações. Por último, utilizaremos métricas semânticas para uma análise mais completa à estrutura concetual. O processo termina após o número de iterações que os especialistas acharem necessário com vista à melhoria constante das propostas de junção de conceitos resultantes do cálculo da similaridade semântica.

### **Passo 1: Normalização (NLP<sup>3</sup>)**

Neste primeiro passo do serviço todos os nomes dos conceitos e suas variantes serão transformados na sua versão linguística base. Esta transformação, independente do contexto, é efetuada recorrendo aos vários mecanismos de processamento de linguagem natural existentes e pretendemos eliminar algumas variações de conceitos que podem influenciar os resultados. Em particular, serão utilizados algoritmos de *Stemming* e *Tokens*, sabendo de antemão que estamos dependentes de um *Dataset* para a língua usada nos nomes dos conceitos. Exemplo: Carros resulta em Carro; Carrinho resulta em Carro.

Do mesmo modo, as relações serão normalizadas de acordo com a sua categoria para nomes mais comuns. De acordo com o tipo de relação (taxonómicas) sub-conceitos (conceitos filhos) podem ser transportados para o conceito mais geral aumentando o conjunto de conceitos que o descrevem.

Após este passo, aumentamos significativamente a qualidade dos resultados, principalmente quando forem aplicadas metodologias sintáticas que visam essencialmente a comparação de texto.

### **Passo 2: Análise de similaridade sintática**

Usando as medidas de comparação de texto, fazemos uma primeira análise recorrendo unicamente aos nomes e variantes dos conceitos. Aqui os resultados indicarão se dois conceitos estão próximos (ou são mesmo iguais) ou distantes sintaticamente. Não podendo garantir que dois conceitos próximos sintaticamente sejam na realidade o mesmo conceito, pode-se no entanto ter em consideração e informar o especialista no sentido da probabilidade de similaridade dos conceitos. Como limitações destas abordagens podemos enumerar as palavras homógrafas, escritas da mesma forma, mas com significados diferentes. Existe também a limitação de não ter em consideração a semântica dos conceitos. Conceitos distantes, ou seja, escritos de forma muito diferente, podem no entanto ser o mesmo conceito. Por este motivo, a nossa abordagem vai recorrer na próxima fase à inclusão de abordagens semânticas para melhorar e encontrar novos conceitos similares.

---

<sup>3</sup> NLP: Processamento de linguagem natural

Neste passo iremos usar algumas das medidas existentes na *framework SimPack*, como por exemplo, *Edit distance* e *Jaro* [Ziegler *et al.* 2006].

### Passo 3: Análise de similaridade semântica

A segunda vertente de análise de similaridade introduz abordagens semânticas para os resultados obtidos no passo 2. Estas abordagens melhoram a qualidade dos resultados uma vez que consideram o significado dos conceitos/relações e não se limitam apenas à construção sintática do nome.

Tendo em conta os recursos do problema já descritos, neste processo de similaridade semântica faremos uso dos seguintes recursos: *Corpus* do projeto; informação definida nos conceitos e relações (variantes, propriedades, definição) e taxonomia (limitado ao domínio em questão pode ou não ser considerada).

**Corpus<sup>4</sup>:** O *corpus* associado ao projeto possui um conjunto de informação (recursos) que ajuda a definir o domínio que os modelos tentam representar. Deste modo, podemos analisar informação relacionada com os conceitos no processo de similaridade. Dentro das abordagens existentes que utilizam o *corpus*, a metodologia seguida é principalmente a procura de padrões no texto. Dados dois conceitos em análise, se eles coocorrerem num dado padrão pode-se inferir algum grau de similaridade entre eles.

Exemplo:

(...) um carro possui rodas (...)

(...) o meio de transporte automóvel possui rodas (...)

Supondo que neste simples excerto os conceitos carro, rodas e automóvel estão presentes nos modelos conceptuais desenvolvidos, é possível identificar um padrão comum relacionado com a propriedade “possui”. Isto porque, em ambos os casos a propriedade tem o mesmo valor: “rodas”. Face a esta partilha de informação, carro e automóvel irão possuir um valor de similaridade mais elevado o que poderá ser um indício de que se trata do mesmo conceito.

**Informação partilhada:** Neste tipo de abordagem faz-se uma análise sobre a informação definida nos conceitos do modelo, no sentido de descobrir informação (propriedades, variantes) partilhada por diferentes conceitos.

---

<sup>4</sup> Corpus: Conjunto de textos que retratam um determinado domínio

Dados dois conceitos em análise, C1 do modelo 1 e C2 do modelo 2, o primeiro passo é analisar a informação comum aos dois. Posteriormente, para aumentar o grau de certeza quanto à similaridade dos dois conceitos, é efetuada uma análise dos seus conceitos diretamente relacionados (filhos). Deste modo, mesmo que da informação de dois conceitos não seja possível indicar que são similares, pela análise em profundidade sobre os conceitos relacionados pode surgir um conjunto de informação comum aos conceitos analisados inicialmente.

**Taxonomia:** O uso da taxonomia é uma etapa à qual se recorre caso as condições assim o permitam (ver Tabela 1 nas desvantagens das medidas baseadas em ontologias). Quando para um dado domínio existir uma taxonomia detalhada e completa a mesma pode ser consultada para calcular similaridades entre conceitos de acordo com abordagens que calculam o tamanho do caminho (*path length*) entre dois conceitos. No nosso problema, é comum não existirem taxonomias suficientemente detalhadas para serem consideradas, porque normalmente são criados modelos de domínios muito específicos. Em suma, as taxonomias poderão ser usadas como um recurso externo para o cálculo da similaridade semântica se nos modelos em questão existirem relacionamentos taxonómicos (IS-A) e a taxonomia definia o mais detalhado possível o domínio do problema. Os conceitos dos modelos conceptuais, ou pelo menos grande parte deles, devem estar presentes na taxonomia para ser possível aplicar as medidas baseadas em taxonomias.

Tendo em conta os diversos tipos de relações que podem existir num modelo conceptual teremos de ter uma ontologia de relações [Sousa *et al.* 2012] para que o cálculo da similaridade semântica não englobe apenas relações taxonómicas (IS-A) e considere também a semântica de outro tipo de relações.

No final, todos os resultados anteriores serão alvo de uma análise para serem filtrados antes da resposta final. É construída uma estrutura que define os conceitos e seus graus de similaridade. Por vezes os resultados de similaridade entre conceitos podem não ser obtidos de forma automática devido às particularidades do domínio em modelação. Por um lado, o ideal seria obter resultados automaticamente, contudo a aplicabilidade destes automatismos está limitada à construção perfeita dos modelos (conceitos bem documentados através de definições, variantes ou outro recurso externo), o que nem sempre acontece. Foi definido um processo iterativo que permite ao utilizador intervir no processo e ajudar a identificar com mais qualidade os conceitos similares. Uma das formas é a definição de âncoras por parte dos especialistas. Uma âncora equivale a um par de conceitos de um modelo, delimitando e definindo um ramo (ou caminho) na árvore de conceitos. A Figura 2 ilustra um exemplo de definição de âncoras nos modelos.



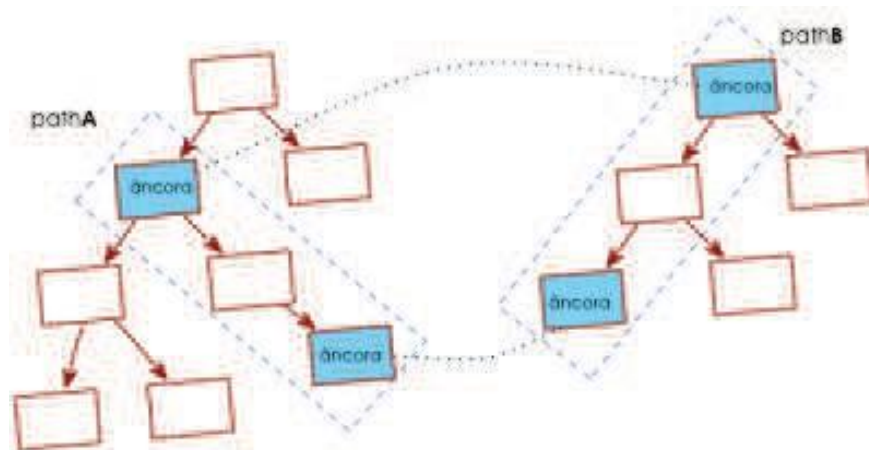


Figura 2 – Definição de âncoras nos modelos

Nesta iteração com os especialistas estamos a focar que a análise de similaridade percorra determinado caminho e aplique as medidas aos conceitos nele existente. Esta definição de âncoras manual permite ajudar o serviço a fazer uma análise mais direcionada enquanto que a ancoragem automática está um pouco dependente da qualidade dos modelos e da informação contida nos mesmos (*corpus* incluído).

### 3. Trabalho relacionado

Embora não exista uma ferramenta ou *framework* que se aplique especificamente ao nosso problema, existem medidas que podem ser integradas na resolução do nosso problema. Partindo das categorias definidas na secção 1 para o cálculo de similaridade semântica, vamos descrever as principais características de cada uma, bem como vantagens e desvantagens.

Dentro das medidas baseadas em ontologias destacam-se: **a) *The shortest path based measure*** - considera a similaridade entre dois conceitos pela distância a que esses conceitos se encontram na taxonomia. Quanto mais próximos estiverem, mais similares são. A distância conceptual entre dois conceitos é proporcional ao número de arestas (uma aresta equivale a uma relação que liga dois conceitos) que os separam na hierarquia [Varelas *et al.* 2005]; **b) *Wu & Palmer's measure*** [Wu & Palmer 1994] - a similaridade de dois conceitos é obtida de acordo com o tamanho das arestas para um conceito específico em comum (um sub-conceito comum); **c) *Leacock & Chodorow's measure*** [Leacock & Chodorow 1998] - medida que calcula a similaridade entre conceitos com base no número de arestas, tendo em conta a profundidade máxima da taxonomia; **d) *Li's measures*** [Li *et al.* 2003] - O cálculo da similaridade é efetuado através da junção das variáveis do tamanho do caminho (número de arestas) e profundidade (número de níveis) numa função não-linear.

Em relação às medidas baseadas no conteúdo (*IC*) temos como referência: **a) *Resnik's measure*** [Resnik 1995] - medida de similaridade semântica baseada no conteúdo.

Dados dois conceitos, o valor da similaridade depende do conteúdo de informação que partilham

na taxonomia; **b) *Lin's measure*** [Lin 2000] - utiliza a informação existente que descreve dois conceitos para atribuir o grau de semelhança entre eles; **c) *Jiang's measure*** [Jiang & Conrath 1997] - calcula a distância semântica entre dois conceitos para obter a similaridade semântica. Após calculado o valor da distância, a similaridade semântica é o inverso desse valor.

***O Tversky's model*** [Tversky 1977] é visto como uma abordagem baseada em características. O cálculo da similaridade semântica contém princípios diferentes das medidas já enunciadas. O autor desta medida afirma que a similaridade não é simétrica, pelo que características entre uma subclasse e a sua superclasse tem uma maior importância para a avaliação da similaridade do que na direção inversa.

Para finalizar, ***Zhou*** [Zhou *et al.* 2008] contribui com uma medida híbrida para o cálculo de similaridade semântica. Esta medida agrupa princípios das medidas anteriormente descritas. Numa primeira fase, tanto o *IC* como o “tamanho do caminho” são considerados. Adicionalmente, é introduzido um peso (parâmetro) definido manualmente que atribui maior relevo aos resultados obtidos pela comparação da informação dos conceitos ou pela sua distância na taxonomia.

A tabela que se segue (Tabela 1) apresenta um breve resumo das principais vantagens e desvantagens dos quatro tipos de categorias para o cálculo da similaridade semântica.



Tabela 1 – Resumo comparativo de cada categoria

<b>Categoria</b>	<b>Princípio</b>	<b>Vantagens</b>	<b>Desvantagens</b>
<b>Baseadas em ontologias</b>	Comprimento do caminho que liga os conceitos e a posição dos conceitos na taxonomia	Aplicação simples	Suporta somente relações IS-A; Obrigatório existir taxonomia; Valores de similaridade iguais para conceitos com a mesma distância e profundidade na taxonomia
<b>Baseadas em IC</b>	Quanto mais informação dois conceitos partilharem mais similares eles são	Utiliza mais informação além do próprio conceito	Pares de conceitos com similaridade igual pela soma do <i>IC</i> de outros pares
<b>Baseadas em características</b>	Conceitos com mais características comuns e com menos características não comuns são mais similares	Características adicionais aos conceitos são consideradas	Elevada complexidade computacional; Não funciona corretamente se o conjunto de características estiver incompleto
<b>Híbridas</b>	Combinam os princípios das categorias anteriores	Distingue com melhor qualidade os pares de conceitos	Necessita de um peso / parâmetro para indicar qual tipo de medida é mais importante, caso contrário pode originar desvios (medida de <i>Zhou</i> )

Para auxiliar o desenvolvimento colaborativo de modelos conceptuais, não existe uma medida que se adegue totalmente. As medidas baseadas em ontologias utilizam como princípio a contagem de arestas e a posição de conceitos numa taxonomia. Deste modo, é feita apenas uma análise a conceitos ligados por relações taxonómicas (IS-A). Outro fator a ter em conta neste tipo de medidas é a necessidade de existir uma taxonomia desenvolvida para os domínios explorados nos modelos conceptuais. Em último caso pode-se consultar uma taxonomia de domínio geral como o *WordNet* mas está limitado ao nível da língua e pode diminuir a qualidade dos resultados.

Utilizando medidas baseadas em *IC* ou em características estamos dependentes da informação disponibilizada pelos especialistas na construção dos modelos. Quanto mais documentados estiverem os conceitos e relações, melhores resultados serão alcançados.

A abordagem descrita neste trabalho vai fazer uso de um conjunto de características das várias categorias para otimizar a qualidade dos resultados tendo em conta as restrições ou condicionantes do problema específico.

#### **4. Conclusão e trabalho futuro**

A similaridade semântica apresenta-se com elevado potencial para resolver problemas relacionados com duplicação e junção de conceitos na construção colaborativa de modelos conceptuais. Devido às condicionantes existentes na conceptualização partilhada, pretendemos aumentar a eficiência e eficácia nas atividades de integração dos modelos retirando grande parte do esforço necessário por parte dos especialistas.

É possível verificar que para o problema em questão é necessário aplicar um conjunto de medidas, que de modo integrado aumentam a qualidade dos resultados como é pretendido. Ambicionamos apresentar uma abordagem que contemple não só a similaridade entre conceitos mas também que não ignore a semântica das relações.

Nos próximos passos iremos colocar em prática o desenvolvimento do serviço e efetuar as primeiras experiências com a sua integração no ConceptME. O objetivo consiste em validar a abordagem e comparar resultados no sentido de aumentar a sua qualidade. A forma de validação irá ser essencialmente baseada no grau de satisfação dos especialistas participantes nas experiências uma vez que não existe uma aplicação do género (aplicada à similaridade semântica de modelos conceptuais) que sirva de comparação.

## 5. Referências

- Costa, L., Sousa, C., Soares, A. L., & Pereira, C. (2012). ConceptME: Gestão colaborativa de modelos conceptuais. In *Capsi*. Guimarães.
- Druckman, D. (2007). Negotiation Models and Applications. In R. Avenhaus & I. W. Zartman (Eds.), *Diplomacy Games SE - 5* (pp. 83–96 LA – English). Springer Berlin Heidelberg.
- Jiang, J., & Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv Preprint Cmp-lg/9709008*, (Rocling X).
- Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An Electronic Lexical Database*, 49(2), 265–283.
- Li, Y., Bandar, Z. A., & McLean, D. (2003). An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Trans. on Knowl. and Data Eng.*, 15(4), 871–882.
- Lin, D. (2000). An Information-Theoretic Definition of Similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 296–304). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Pereira, C. (2010). *A organização da informação e conhecimento em redes colaborativas como um processo de construção social do significado: uma teoria e um método prático*. PhD. Faculdade de Engenharia Da Universidade Do. FEUP.
- Pereira, C., Sousa, C., & Lucas Soares, A. (2012). Supporting conceptualisation processes in collaborative networks: a case study on an R&D project. *International Journal of Computer Integrated Manufacturing*, 1–21.
- Petrakis, E., Varelak, G., Hliaoutakis, A., & Raftopoulou, P. (2006). Design and evaluation of semantic similarity measures for concepts stemming from the same or different ontologies. *4th Workshop on Multimedia Semantics (WMS'06)*, 44–52.
- Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1* (pp. 448–453). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Sánchez, D., & Batet, M. (2012). A semantic similarity method based on information content exploiting multiple ontologies. *Expert Systems with Applications*, 40, 1393–1399.
- Sousa, C., Soares, A. L., Pereira, C., & Costa, R. (2012). Supporting the identification of conceptual relations in semi-formal ontology development. Istanbul, Turkey.

- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(2), 327–352.
- Varelas, G., Voutsakis, E., Petrakis, E. G. M., Milios, E. E., & Raftopoulou, P. (2005). Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web. In *In: 7 Th ACM Intern. Workshop on Web Information and Data Management (WIDM 2005* (pp. 10–16). ACM Press.
- Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics* - (pp. 133–138). Morristown, NJ, USA: Association for Computational Linguistics.
- Zhou, Z., Wang, Y., & Gu, J. (2008). New model of semantic similarity measuring in wordnet. In *Intelligent System and Knowledge Engineering, 2008. ISKE 2008. 3rd International Conference On* (Vol. 1, pp. 256–261).
- Ziegler, P., Kiefer, C., & Sturm, C. (2006). Detecting similarities in ontologies with the SOQA-SimPack toolkit. *Advances in Database*.