Manipulação Semântica de Textos
Os Projetos Wordnet e LSA

Sergio C. Navega
Intelliwise AI Research
snavega@intelliwise.com
Publicado no Infoimagem 2004
Agosto de 2004

Resumo

Neste artigo serão apresentados dois sistemas distintos para a manipulação semântica de textos. A principal finalidade deste texto é fazer o leitor conhecer um pouco sobre essas interessantes idéias. No processo LSA, o fundamento é estatístico. Através do uso de algumas operações de álgebra linear são obtidas representações truncadas das ocorrências de palavras em um grande grupo de documentos, o que permite um acesso à semântica das palavras, com várias aplicações como recuperação de informação, análise de coerência de textos, classificação de textos e outras. O projeto WordNet parte de pressupostos diferentes. O WordNet é uma base de informações léxicas com diversas características ligando as palavras entre si. Isto permite a utilização do sistema também para a recuperação de informação, categorização de textos, inferência e coerência de textos e até mesmo conectando-se com a Web Semântica.

Introdução

Nossa civilização tem armazenado um gigantesco repositório de informações textuais no formato de livros, revistas, manuais e documentos diversos. Boa parte do progresso científico, cultural e tecnológico que experimentamos nos últimos séculos deve-se a esse repositório. Através dele, somos capazes de pinçar um pouco da sabedoria e do esforço de pensamento de milhões de nossos antepassados. Contudo, até recentemente esses repositórios estavam essencialmente sob a forma impressa. Hoje temos um número exponencialmente crescente de informações sendo armazenadas e disponibilizadas em formato eletrônico, tanto em mídias digitais quanto via Internet. Alguns pensadores consideram que, por causa desses novos formatos, nossa sociedade poderá ter um incremento de produtividade intelectual comparável ao da disseminação de original de material impresso. Mas para que isso se torne realidade, é preciso superar um grande desafio, pois *maior quantidade* de informação não significa necessariamente *melhor qualidade* de informação. Esta é a raiz do problema que é preciso resolver: o que podemos fazer para selecionar, categorizar e pesquisar o gigantesco volume de material textual de que dispomos?

Neste artigo veremos um processo (LSA) que pode nos auxiliar nessa empreitada. Veremos também alguns detalhes de um projeto interessante (Wordnet) que se dedicou à construção de um dicionário semântico.

Por Que Pesquisar Em Textos É Tão Difícil?

É de certa forma surpreendente a dificuldade que temos para selecionar material textual relevante a partir de algumas palavras chaves. Em termos computacionais, este problema já foi resolvido há tempos, mas os resultados não parecem ser o que desejamos. Quando se fala de mecanismos de busca como Google ou Yahoo tem-se uma fonte de informação preciosa e rápida, capaz de colocar na ponta de nossos dedos milhares de referências sobre o assunto que desejamos. Mesmo assim, é comum gastarmos muito tempo selecionando aquilo que realmente nos interessa. Por isso, essa forma de acesso não é capaz de potencializar o uso da informação disponível eletronicamente. O problema típico que enfrentamos está relacionado à semântica das palavras, ou seja, *ao seu significado*. Na verdade, há como argumentar que somente agentes inteligentes podem desenvolver uma real semântica do mundo natural[1]. Contudo, deixando de lado essa questão filosófica podemos encontrar algumas aproximações muito úteis.

Começamos por identificar o problema mais de perto: imagine que você deseja fazer uma pesquisa sobre a crise do petróleo. Sabemos que qualquer documento que fale em "crise da gasolina" deveria lhe interessar. Mas isso *não é retornado* pelos mecanismos de busca tradicionais, pois 'gasolina' não é uma palavra formada pelo mesmo conjunto de caracteres de 'petróleo'. No entanto, essas duas palavras estão muito próximas em termos de semântica. O quadro abaixo resume três das potenciais dificuldades dos sistemas tradicionais de busca:

Descrição	Nome	Sistemas Tradicionais	Exemplos
Palavras que tem mais de um significado	Polisemia	Afetam a precisão das pesquisas	morsa, roda, cartucho
Várias palavras que significam a mesma coisa	Sinonimia	Baixo número de retornos de uma pesquisa	grana, dinheiro, valores, ações, recursos
Palavras fortemente associadas a outras	Associação	Não retornam textos com palavras associadas	gasolina, petróleo, Iraque, Golfo Pérsico, OPEP

A polisemia afeta a precisão das pesquisas, retornando muitos documentos que não têm relação com a pesquisa desejada. Já no caso da sinonimia, são retornados poucos documentos, mesmo que existam muitos que poderiam ser retornados por causa de sinônimos. Mas o pior caso é justamente o de *palavras associadas*: nenhum documento é retornado, embora sejam claramente relevantes. É preciso algo mais poderoso do que a simples busca literal por palavras chaves.

Esboço de Uma Solução

O problema de pesquisar textos através de noções semânticas ganhou uma importante contribuição com a introdução de uma linha de técnicas conhecida por LSI (Latent Semantic Indexing). A idéia central desses algoritmos é processar os documentos e extrair deles uma *representação reduzida* que facilite a busca. Contudo, o segredo dessa idéia está relacionado a como essa representação é obtida. Utiliza-se um processo que obtém um agrupamento das palavras de acordo com os seus "sentidos conceituais" mais apropriados. Nos processos tradicionais, há uma preocupação em obter a indexação através da posição de cada palavra. No processo LSI, as palavras são consideradas dentro do *contexto* em que estão inseridas, ou seja, o

método captura a significação estatística da palavra *em relação às palavras que a circundam*. O quadro abaixo apresenta um exemplo de consulta utilizando este método:

Palavra da pesquisa	Retorno de alta relevância	Palavras com baixa relevância
médico	Hospital Enfermeira Tratamento Paciente Medicamento	Madeira Asfalto Árvore
carro	Automóvel Motorista Veículo Chassis	Elefante Computador Telefone

As palavras da coluna do meio são retornadas como relevantes, ou seja, farão os documentos a elas associados serem retornados por causa da pesquisa da palavra na coluna da esquerda. A última coluna mostra alguns exemplos de palavras cuja significação é muito baixa, ou seja, não irão provocar o retorno de documentos associados. A tabela abaixo apresenta uma comparação do potencial de recuperação das palavras chaves e do método LSA.

Palavras	Porcentual de Recuperação Mecanismos Tradicionais	Porcentual Recuperação Mecanismo LSA
doutor – doutor	1.0	1.0
doutor – médico	0	0.8
doutor – cirurgião	0	0.7

O processo LSA é tão poderoso que possibilita a recuperação de um documento que contenha a expressão "o raio das esferas" a partir da digitação de "o diâmetro do círculo" (porcentual de recuperação = 55%).

Como Funciona o LSA

LSA (Latent Semantic Analysis) é o nome do processo baseado em LSI que providencia os resultados que vimos no quadro anterior. A técnica tem sua origem associada a uma área da matemática conhecida como Análise de Fatores (Factor Analysis)[2]. Descreveremos a seguir, de forma muito simplificada, todo o processo.

Começamos com uma relação de documentos que desejamos indexar. Esses documentos podem ser tão curtos quanto um parágrafo ou tão grandes quanto um ensaio. Textos pequenos permitem uma maior precisão, mas requerem mais recursos de processamento. Desse grupo de textos extraímos uma gigantesca matriz T com esta constituição:

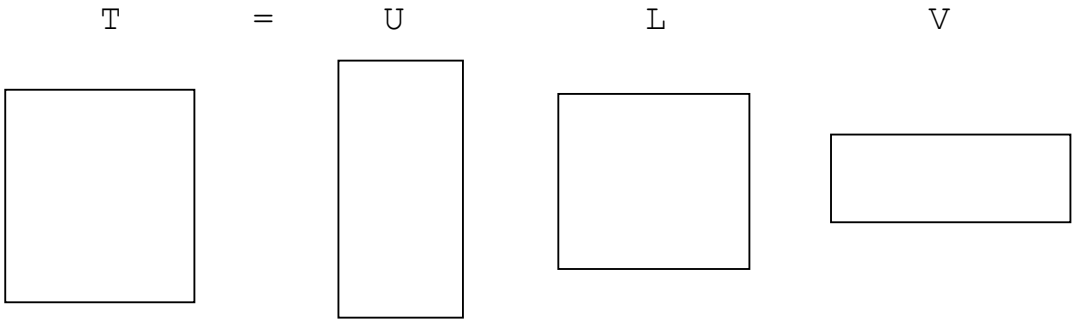
	D1	D2	D3	.....
Palavra1	0	1	0	.....
Palavra2	3	0	0	.....
Palavra3	1	4	0	.....
.....	.....	.....	.....	.....

As linhas dessa tabela representam cada uma das palavras que serão indexadas[3]. As colunas representam cada um dos documentos e a intersecção entre linhas e colunas, o número de vezes que a palavra ocorreu naquele documento específico. Assim, a Palavra1 não ocorreu nenhuma vez no documento D1 e ocorreu 1 vez no documento D2. A Palavra2 ocorreu 3 vezes em D1 e nenhuma vez em D2 e assim por diante. Após a montagem dessa matriz, são feitos alguns ajustes que não tem importância para nossa explicação aqui[4].

Após essas transformações, a matriz de termos atravessará o cerne do processo: a decomposição em valores singulares (ou SVD-Singular Value Decomposition). Essa técnica tem origem em álgebra linear e o resultado é a transformação da matriz de termos original em três outras matrizes, U, L e V. A multiplicação dessas matrizes reconstitui a matriz original:

$$\mathbf{T = U \,.\, L \,.\, V}$$

O diagrama abaixo mostra essa operação ressaltando o formato de matriz:



Matriz Original (m x n)	Vetores de Termos (m x r)	Valores Singulares (r x r)	Vetores de Documentos (r x m)

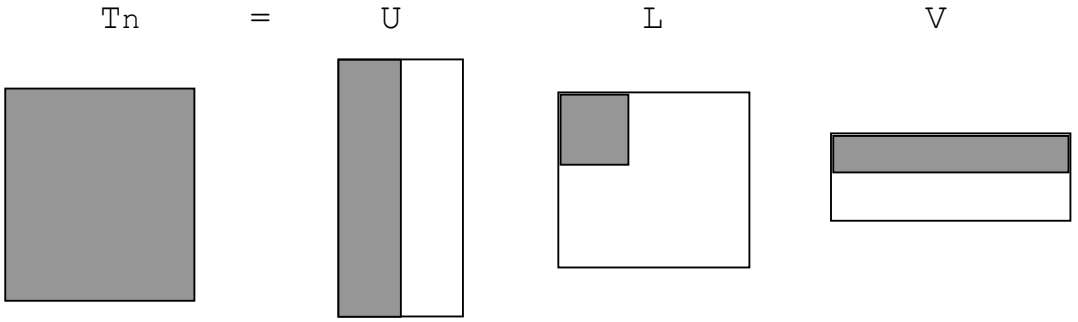
onde (m x n) é a dimensão[5] da matriz original e r é a ordem de T (tipicamente o mínimo entre m e n). A matriz L (valores singulares) é *diagonal*, ou seja, apenas a diagonal contém valores significativos. A dimensão dessas matrizes é usualmente bastante grande, já que o número de documentos (e também de palavras indexadas) também é muito grande. Contudo, o ponto essencial de todo este processo ainda está por ser dado. Mas antes de explica-lo é necessário introduzir uma noção importante.

O Efeito da Redução de Informação

Quem trabalha na área de TI de qualquer empresa tem claro uma visão bastante disseminada: quanto mais informação melhor. Dito em outras palavras, é consenso achar que qualquer processo no qual há *perda* de informações, estaremos também reduzindo o valor. Essa visão é apropriada quando falamos de bancos de dados ou em transações. Mas ela *não é adequada* quando estamos procurando formas de generalização[6]. Para generalizar precisamos concentrar nossa atenção apenas nos fatores que são estatisticamente relevantes.

Por essa razão, o passo seguinte de nosso processo LSA é uma restrição (redução) no tamanho da matriz de valores singulares. Em vez de mantê-la com o tamanho (r x r), fazemos

uma redução para um tamanho (k x k), onde k é um número muito menor do que r. Todos os valores além de k serão transformados em zero. Isto irá transformar nossa matriz de valores singulares em uma versão "mais enxuta", contendo apenas poucos elementos. Por isso, nossa fórmula original fica assim transformada:



Matriz Original (m x n)	Vetores de Termos (m x r)	Valores Singulares (r x r)	Vetores de Documentos (r x m)

Apenas as áreas acinzentadas contêm valores significativos. Esta "redução" de áreas significativas irá montar uma *nova* matriz de termos T (chamaremos de Tn) que possuirá elementos bastante diferentes da matriz original. Esta operação de redução chama-se "SVD truncado" e possibilita uma redução da influência individual dos termos preservando os *padrões principais de uso* desses termos. Cada coluna da nova matriz Tn será uma descrição (um vetor) que representará diferentes padrões de uso das palavras mais relevantes desse documento. Isto permite que seja possível *comparar* documentos de acordo com um critério estatisticamente interessante[7].

Elaborando Consultas

Após o processamento SVD conforme descrito, chega a hora de permitir ao usuário que estabeleça uma consulta. O usuário digitará uma série de palavras e espera que o sistema possa lhe entregar os documentos que possuam maior relevância semântica com essas palavras. A consulta do usuário (a lista de palavras) forma um novo "mini-documento" que passa pelo mesmo processo descrito anteriormente. O resultado é um vetor descrevendo (tal qual a indexação anterior) o padrão de uso das palavras. Agora, basta comparar esse vetor com os da matriz Tn (na verdade, o processo é um pouco mais complexo). Dessa comparação obteremos os vetores que mais se aproximam do vetor da consulta. Classificamos esses vetores em ordem de similaridade (os mais similares ficarão no topo da lista) e apresentamos ao usuário apenas os documentos associados a esses primeiros da lista.

Quais as Vantagens do Processo?

O processo LSA possibilita a recuperação de informação textual a partir de textos semanticamente associados às palavras usadas na consulta. Isto só é possível porque existe uma significativa redução de informação, com a manutenção apenas daquilo que realmente importa. As matrizes de termos T têm usualmente uma dimensão muito grande. Como exemplo, se você

tem 12.000 documentos dos quais separa 40.000 palavras, a matriz T terá 480 milhões de elementos. Contudo, após passar pelo processo SVD truncado ficamos tipicamente com uma matriz bem mais modesta:

As aplicações deste processo são inúmeras:

a) Recuperação de informação
   Possibilita pesquisas por noções semânticas, e não apenas por palavras-chave.
b) Feedback de relevância
   Permite que o usuário refine progressivamente sua pesquisa.
c) Filtragem de informação
   Exibe conjuntos de textos que passem por um "filtro" semântico.
d) Recuperação em outras linguagens
   A pesquisa LSA permite a comparação de textos mesmo de línguas diferentes
e) Seleção e avaliação de textos educacionais
   É possível estabelecer critérios de avaliação de ensaios escritos por estudantes.
f) Roteamento de informação
   Permite encaminhar textos (emails ou noticiário eletrônico) para as pessoas certas.

O processo LSA tem importantes implicações para a ciência da computação e também para a filosofia da mente (veja, por exemplo, o trabalho de Landauer e Dumais 1997).

O Projeto WordNet

Dicionários eletrônicos são ferramentas que não parecem ter evoluído muito desde sua introdução há algumas décadas atrás. Mas a partir de 1985, pesquisadores da Universidade de Princeton, liderados por George Miller, cogitaram em utilizar algumas hipóteses lingüísticas para formatar um novo tipo de dicionário. Miller e sua equipe seguiram a idéia de que os componentes léxicos[8] de uma linguagem poderiam ser isolados do restante (gramática, por exemplo). Assim nasceu o WordNet, uma Base de Dados Léxica que contém informações sobre palavras, palavras compostas, verbos, frases idiomáticas, relações hierárquicas entre palavras e outras propriedades. Essa base de dados possui, por exemplo, informações sobre sinônimos (palavras diferentes que significam a mesma coisa), hiperônimos/hipônimos (a derivação hierárquica das palavras), merônimos (as "partes" associadas ao sentido de uma palavra), etc.

Com o WordNet diversas aplicações tornam-se possíveis, não apenas na recuperação de informação mas também na análise do significado de frases. Como exemplo, há sistemas baseados em WordNet (veja Harabagiu e Moldovan 1998) que recebem frases e deduzem certas propriedades:

João estava faminto
Ele abriu a geladeira

Se perguntarmos “Por que João abriu a geladeira”?, um sistema baseado em WordNet pode responder “Para se alimentar”. Este resultado é obtido através de inferências feitas sobre a base léxica, onde se combinam duas ou mais relações semânticas[9].

Contendo mais de 138 mil palavras inglesas ligadas através de centenas de milhares de formas diferentes, WordNet potencializa o uso da informação estéril contida nos dicionários tradicionais. Os quadros a seguir apresentam algumas informações retornadas pelo sistema WordNet em relação à palavra “automobile”.

Hipernimia da palavra “automobile” **“automobile” é um tipo de…**

Sense 1
car, auto, automobile, machine, motorcar -- (4-wheeled motor vehicle; usually propelled by an internal combustion engine;
"he needs a car to get to work")
=> motor vehicle, automotive vehicle -- (a self-propelled wheeled vehicle that does not run on rails)
=> self-propelled vehicle -- (a wheeled vehicle that carries in itself a means of propulsion)
=> wheeled vehicle -- (a vehicle that moves on wheels and usually has a container for transporting things or people;
"the oldest known wheeled vehicles were found in Sumer and Syria and date from around 3500 BC")
=> vehicle -- (a conveyance that transports people or objects)
=> conveyance, transport -- (something that serves as a means of transportation)
=> instrumentality, instrumentation -- (an artifact (or system of artifacts) that is instrumental in accomplishing some end)
=> artifact, artefact -- (a man-made object taken as a whole)
=> object, physical object -- (a tangible and visible entity; an entity that can cast a shadow; "it was full of rackets, balls and other objects")
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
=> whole, whole thing, unit -- (an assemblage of parts that is regarded as a single entity; "how big is that part compared to the whole?"; "the team is a unit")
=> object, physical object -- (a tangible and visible entity; an entity that can cast a shadow; "it was full of rackets, balls and other objects")
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
=> container -- (any object that can be used to hold things (especially a large metal boxlike object of standardized dimensions that can be loaded from one form of transport to another))
=> instrumentality, instrumentation -- (an artifact (or system of artifacts) that is instrumental in accomplishing some end)
=> artifact, artefact -- (a man-made object taken as a whole)
=> object, physical object -- (a tangible and visible entity; an entity that can cast a shadow; "it was full of rackets, balls and other objects")
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
=> whole, whole thing, unit -- (an assemblage of parts that is regarded as a single entity; "how big is that part compared to the whole?"; "the team is a unit")
=> object, physical object -- (a tangible and visible entity; an entity that can cast a shadow; "it was full of rackets, balls and other objects")
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Hiponimia da palavra “automobile” **….. é um tipo de “automobile”**

Sense 1
car, auto, automobile, machine, motorcar -- (4-wheeled motor vehicle; usually propelled by an internal combustion engine; "he needs a car to get to work")
=> ambulance -- (a vehicle that takes people to and from hospitals)
=> beach wagon, station wagon, wagon, beach waggon, station waggon, waggon -- (a car that has a long body and rear door with space behind rear seat)
=> bus, jalopy, heap -- (a car that is old and unreliable; "the fenders had fallen off that old bus")
=> cab, hack, taxi, taxicab -- (a car driven by a person whose job is to take passengers where they want to go in exchange for money)
=> compact, compact car -- (a small and economical car)
=> convertible -- (a car that has top that can be folded or removed)
=> coupe -- (a car with two doors and front seats and a luggage compartment)
=> cruiser, police cruiser, patrol car, police car, prowl car, squad car -- (a car in which policemen cruise the streets; equipped with radiotelephonic communications to headquarters)
=> electric, electric automobile, electric car -- (a car that is powered by electricity)
=> gas guzzler -- (a car with relatively low fuel efficiency)
=> hardtop -- (a car that resembles a convertible but has a fixed rigid top)
=> hatchback -- (a car having a hatchback door)
=> horseless carriage -- (an early term for an automobile; "when automobiles first replaced horse-drawn carriages they were called horseless carriages")
=> hot rod, hot-rod -- (a car modified to increase its speed and acceleration)
=> jeep, landrover -- (a car suitable for traveling over rough terrain)
=> limousine, limo -- (large luxurious car; usually driven by a chauffeur)
=> loaner -- (a car that is lent as a replacement for one that is under repair)
=> minicar -- (a car that is even smaller than a subcompact car)
=> minivan -- (a small box-shaped passenger van; usually has removable seats; used as a family car)
=> Model T -- (the first widely available automobile powered by a gasoline engine; mass-produced by Henry Ford from 1908 to 1927)
=> pace car -- (a high-performance car that leads a parade of competing cars through the pace lap and then pulls off the course)
=> racer, race car, racing car -- (a fast car that competes in races)
=> roadster, runabout, two-seater -- (an open automobile having a front seat and a rumble seat)
=> sedan -- (a car that is closed and that has front and rear seats and two or four doors)
=> sports car, sport car -- (a small low car with a high-powered engine; usually seats two persons)
=> sport utility, sport utility vehicle, S.U.V. -- (a high-performance four-wheel drive car built on a truck chassis)
=> Stanley Steamer -- (a steam-powered automobile)
=> stock car -- (a car kept in dealers' stock for regular sales)
=> subcompact, subcompact car -- (a car smaller than a compact car)
=> touring car, phaeton, tourer -- (large open car seating four with folding top)

Meronimia da palavra “automobile” **“automobile” é constituído de....**

Sense 1
car, auto, automobile, machine, motorcar -- (4-wheeled motor vehicle; usually propelled by an internal combustion engine; "he needs a car to get to work")
- HAS PART: accelerator, accelerator pedal, gas pedal, gas, throttle, gun -- (a pedal that controls the throttle valve; "he stepped on the gas")
- HAS PART: air bag -- (a safety restraint in an automobile; the bag inflates on collision and prevents the driver or passenger from being thrown forward)
- HAS PART: auto accessory -- (an accessory for an automobile)
- HAS PART: automobile engine -- (the engine that propels an automobile)
- HAS PART: automobile horn, car horn, motor horn, horn, hooter -- (a device on an automobile for making a warning noise)
- HAS PART: buffer, fender -- (a cushion-like device that reduces shock due to contact)
- HAS PART: bumper -- (a mechanical device consisting of bars at either end of a vehicle to absorb shock and prevent serious damage)
- HAS PART: car door -- (the door of a car)
- HAS PART: car mirror -- (a mirror that the driver of a car can use)
- HAS PART: car seat -- (a seat in a car)
- HAS PART: car window -- (a window in a car)
- HAS PART: fender, wing -- (a barrier that surrounds the wheels of a vehicle to block splashing water or mud; "in England they call a fender a wing")
- HAS PART: first gear, first, low gear, low -- (the lowest forward gear ratio in the gear box of a motor vehicle; used to start a car moving)
- HAS PART: floorboard -- (the floor of an automobile)
- HAS PART: gasoline engine -- (an internal-combustion engine that burns gasoline; most automobiles are driven by gasoline engines)
- HAS PART: glove compartment -- (compartment on the dashboard of a car)
- HAS PART: grille, radiator grille -- (grating that admits cooling air to car's radiator)
- HAS PART: high gear, high -- (a forward gear with a gear ratio giving high vehicle velocity for a given engine speed)
- HAS PART: hood, bonnet, cowl, cowling -- (protective covering consisting of a metal part that covers the engine; "there are powerful engines under the hoods of new cars"; "the mechanic removed the cowling in order to repair the engine")
- HAS PART: luggage compartment, automobile trunk, trunk -- (compartment in an automobile that carries luggage or shopping or tools; "he put his golf bag in the trunk")
- HAS PART: rear window -- (car window that allows vision out of the back of the car)
- HAS PART: reverse -- (the gears by which the motion of a machine can be reversed)
- HAS PART: roof -- (protective covering on top of a motor vehicle)
- HAS PART: running board -- (a narrow footboard serving as a step beneath the doors of some old cars)
- HAS PART: stabilizer bar, anti-sway bar -- (a rigid metal bar between the front suspensions and between the rear suspensions of cars and trucks; serves to stabilize the chassis)
- HAS PART: sunroof, sunshine-roof -- (an automobile roof having a sliding or raisable panel; "`sunshine-roof' is a British term for `sunroof'")
- HAS PART: tail fin, tailfin, fin -- (one of a pair of decorations projecting above the rear fenders of an automobile)
- HAS PART: third gear, third -- (the third from the lowest forward ratio gear in the gear box of a motor vehicle; "you shouldn't try to start in third gear")
- HAS PART: window -- (a transparent opening in a vehicle that allow vision out of the sides or back; usually is capable of being opened)

Diversas aplicações acadêmicas têm sido criadas com o uso do WordNet. Entre elas citamos Recuperação de informação baseada em bases de conhecimento (Richardson & Smeaton 1995), recuperação de imagens através de descrições (Smeaton & Quigley 1996) e categorização de textos (Gomes-Hidalgo & Rodriguez 1997; Rodriguez, Gómez-Hidalgo & Diaz-Agudo 1997). Mas com a iniciativa da Web Semântica e com o uso de conversões do WordNet para RDF (veja links recomendados) as possibilidades de utilização desta iniciativa serão multiplicadas substancialmente.

Referências

Berry, Michael W.; Dumais, Susan; O'Brien, G. W. (1994) Using Linear Álgebra for Intelligent Information Retrieval. University of Tennessee, CS Department, CS-94-270.

Berry, Michael W.; Dumais, Susan; Shippy, A. T. (1995) A Case Study of Latent Semantic Indexing. University of Tennessee, CS Department, CS-95-271.

Berry, Michael W.; Letsche, Todd A. (1997) Large-Scale Information Retrieval with Latent Semantic Indexing. Information Sciences-Applications, 1997.

Deerwester, Scott; Dumais, Susan T.; Furnas, George W.; Landauer, Thomas K.; Harshman, Richard (1990) Indexing by Latent Semantic Analysis. Journal of the American Society of Information Science.

Dumais, Susan (1996) Combining Evidence for Effective Information Filtering.

Dumais, Susan; Landauer, Thomas; Littman, Michael (1996) Automatic Cross-Linguistic Information Retrieval Using Latent Semantic Indexing.

Gómez-Hidalgo, José María; Rodriguez, Manuel de Buenaga (1997) Integrating a Lexical Database and a Training Collection for Text Categorization. ACL/EACL Workshop on Automatic Extraction of Lexical Semantic Resources for NL Applications, 1997.

Harabagiu, Sanda M.; Moldovan, Dan I. (1998) Knowledge Processing on an Extended WordNet. In: Fellbaum, Christiane (ed) WordNet An Electronic Lexical Database. MIT Press.

Hearst, Marti A. (1992) Automatic Acquisition of Hyponyms from Large Text Corpora. 14th International Conference on Computational Linguistics.

Landauer, Thomas; Littman, Michael (1995) A Statistical Method for Language-Independent Representation of the Topical Content of Text Segments. 6th Annual Conference UW Centre for the New Oxford English Dictionary and Text Research.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.

Navega, Sergio C. (2000) Inteligência Artificial, Educação de Crianças e o Cérebro Humano. Leopoldianum, Revista de Estudos de Comunicações da Universidade de Santos, Ano 25, No.72, Fev. 2000, pp 87-102. Disponível em http://www.intelliwise.com/reports/p4port.htm

Navega, Sergio C. (2002) Projeto CYC: Confundindo Inteligência com Conhecimento. KMBrasil 2002, 3o. Workshop Brasileiro de Inteligência Competitiva. Disponível em http://www.intelliwise.com/reports/kmbscn.htm

Richardson, R.; Smeaton, A. F. (1995) Using WordNet in a Knowledge-Based Approach to Information Retrieval. University of Dublin technical report CA-0395.

Rodriguez, Manuel de Buenaga; Gómez-Hidalgo, José María; Díaz-Agudo, Belén (1997) Using WordNet to Complement Training Information in Text Categorization. Recent Advances in Natural Language Processing, 1997.

Salton, G; McGill, M. J. (1983) Introduction to Modern Information Retrieval. McGraw-Hill.

Shannon, Claude (1948) A Mathematical Theory of Communication. Bell Systems Technical Journal, 27:379-423, 623-656, 1948.

Smeaton, A. F; Quigley, I. (1996) Experiments on Using Semantic Distances Between Words in Image Caption Retrieval. University of Dublin technical report CA-0196.

Voorhees, Ellen M. (1998) Using WordNet for Text Retrieval. In: Fellbaum, Christiane (ed) WordNet An Electronic Lexical Database. MIT Press.

Links Recomendados

WordNet, A Lexical Database for the English Language
http://www.cogsci.princeton.edu/~wn/

WordNet Bibliography
http://engr.smu.edu/~rada/wnb/

Biblioteca SVDPACK
http://www.netlib.org/svdpack/

Latent Semantic Analysis Web Site
http://lsa.colorado.edu/

Readings in Latent Semantic Analysis
http://www.upmf-grenoble.fr/sciedu/blemaire/lsa.html

A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge
http://lsa.colorado.edu/papers/plato/plato.annote.html

Semantic Web
http://en.wikipedia.org/wiki/Semantic_Web

A RDF Representation of WordNet
http://www.semanticweb.org/library/

Notas

[1] Este tópico está melhor apresentado em Navega (2000). Basicamente, para se desenvolver uma semântica do mundo natural é preciso, além de inteligência, ter um íntimo contato sensório com o mundo. Além disso, vários conceitos complexos só podem aparecer em organismos que *interajam* com esse mundo, ou seja, atuem e percebam os resultados dessas ações.

[2] Há uma família de técnicas matemáticas conhecidas como PCA (Principal Component Analysis), ICA (Independent Component Analysis) e Eingenvector Decomposition, entre outras, cuja idéia básica é abstrair de uma sequência de dados os fatores ou componentes *estatisticamente relevantes*. Em todas essas noções existe a idéia básica de reduzir a complexidade (redução de dimensionalidade) dos dados deixando os componentes que parecem ser realmente relevantes. O leitor interessado em mais informações sobre o funcionamento dessas idéias deve seguir as referências (principalmente os textos de Dumais, Berry e Landauer).

[3] Normalmente, são eliminadas da indexação as palavras de uma letra só, as que ocorrem em apenas um documento e as palavras que estão em uma "lista de parada", ou seja, aquelas que não carregam nenhum significado importante (exemplo: quem, porque, todo, acerca, outro, antes, além, nada, fora, alguém, através, enquanto, etc).

[4] Esses ajustes tipicamente são efetuados para transformar as freqüências de ocorrência em uma modalidade não-linear (tipicamente logarítmica log(freq)). Além disso, é feito um ajuste de cada termo para compensar a "raridade" da palavra no conjunto de documentos (tipicamente uma medida do inverso da freqüência de ocorrência nos documentos ou um fator que leva em conta efeitos relacionados à entropia).

[5] Uma matriz (3 x 2) significa um grupo de 3 linhas e 2 colunas.

[6] A importância dos processos de generalização não deve ser subestimada. Muito daquilo que aprendemos acerca do ambiente que nos cerca tem que ser generalizado, e generalização envolve necessariamente alguma forma de "perda" de informação. Na verdade, devemos falar não em perda, mas sim em *descarte*: fica-se apenas com aquilo que tem relevância para o nosso comportamento. Para mais detalhes sobre estes assuntos, veja Navega (2000) e Navega (2002).

[7] O critério mais utilizado é o coseno do ângulo formado pelo vetor em um espaço k-dimensional. Vetores "semelhantes" (ou seja, com padrões de significado próximos) terão cosenos similares.

[8] Léxico de uma linguagem é uma coleção de palavras junto com seus sentidos usuais. Poderíamos entender o léxico como uma espécie de dicionário onde cada palavra é analisada através de seu(s) sentido(s) e também de várias propriedades, como sinonímia, hipernimia, meronimia, etc.

[9] A explicação de como esta técnica funciona foge do escopo deste artigo introdutório. Vale dizer que as idéias desta técnica baseiam-se em um procedimento conhecido como "marker passing" que simula a propagação de associações semânticas entre unidades léxicas da base WordNet, algo que tem correspondentes com o que ocorre no cérebro humano.