UNIVERSIDADE FEDERAL DE SANTA CATARINA

Construindo Aplicações na Web Semântica
Buscas Semânticas

Renato Fileto
fileto@inf.ufsc.br

Programa de Pós-graduação em Ciência da Computação – PPGCC
Departamento de Informática e Estatística – INE
Centro Tecnológico – CTC

Conteúdo

- **Motivação**
- **Definição do problema**
- **Abordagens e ferramentas**
- **Conclusões**

Recuperação da Informação

- **Medidas da eficácia das soluções**

precisão = $\frac{\text{nro. de docs relevantes recuperados}}{\text{nro. de docs recuperados}}$

cobertura = $\frac{\text{nro. de docs relevantes recuperados}}{\text{nro. de docs relevantes}}$

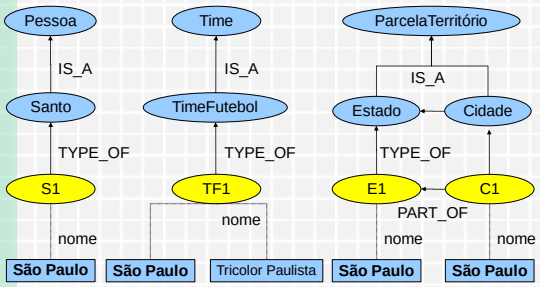
Problemas da RI tradicional

- Usa somente informação léxica e sintática
- Traz todos os resultados relacionados sintaticamente com uma palavra-chave, sem organização quanto aos possíveis significados
- Não recupera resultados cujo conteúdo ou metadados referencie outras palavras com significado parecido ou relacionado
- Menospreza as características dos usuários quanto ao significado que atribuem a determinados termos

Buscas semânticas

- Consideram os significados (das consultas e do conteúdos ou metadados), procurando:
  - evitar, identificar e/ou tratar ambigüidades,
  - expandir semanticamente as buscas para significados relacionados,
  - orientar o usuário no refinamento das suas consultas,
  - apresentar os resultados classificados segundo os significados a eles atribuídos
  - ordenar os resultados segundo a relevância estimada dos diferentes significados para a consulta e o usuário que a estipulou,
  - Tornar mais rica, agradável e efetiva a interação do usuário com o sistema de busca.

Trecho de ontologia 1



Classificação de abordagens para buscas semânticas (Mangold, 2007)

- Arquitetura
- Acoplamento entre ontologias e documentos
- Transparência
- Contexto do usuário
- Modificação de consultas
- Estrutura das ontologias
- Tecnologia para representar as ontologias
- Desempenho, escalabilidade e distribuição
- Adaptabilidade
- Ranking

Arquiteturas para buscas semânticas

- **Máquina de busca *stand_alone*:** armazena metadados sobre os documentos em índices com os quais avalia as consultas
- **Meta-máquina de busca:** distribui as consultas para outras máquinas de busca e combina os resultados

Acoplamento ontologias-documentos

- **Acoplamento forte:** os metadados dos documentos referem-se explicitamente aos conceitos de uma ontologia e vice-versa (requer anotação semântica)
- **Acoplamento fraco:** os documentos são independentes de quaisquer ontologias disponíveis (pode restringir a capacidade de busca)

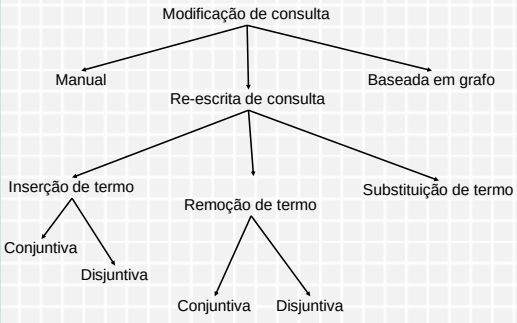
Contexto do usuário

- ***Learning*:** extraído na interações do usuário com o sistema (histórico das consultas e refinamentos de consultas)
- ***Hard-coded*:** as consultas são organizadas em categorias que definem a informação solicitada (e.g., "*location of*", "*resources for*" ...)

Modificação de consultas

- **Manual:** o sistema retorna um trecho apropriado da ontologia junto com a resposta, permitindo ao usuário navegar na ontologia e reformular a consulta pela adição e remoção de termos
- **Re-escrita automática:** a consulta é otimizada automaticamente pelo sistema
- **Baseada em grafo:** considera conceitos e documentos como nodos de um grafo, efetuando propagação da consulta a partir dos nodos correspondentes a termos de busca

Modificação de consultas



Estrutura da ontologia

- **Propriedades anônimas:** as relações entre conceitos não são nomeadas, indicando apenas que eles compartilham o mesmo contexto
- **Propriedades padronizadas:** sinônimia, hipernímia, meronímia, negação, etc.
- **Propriedades específicas de domínio:** além das propriedades padrão, considera propriedades específicas de um domínio (e.g., "tipo de câmera" em sistema para a área de fotografia)

Tecnologia da ontologia

- F-Logic
- RDF
- DAML+OIL
- Frames
- OWL
- :

Comparação de abordagens

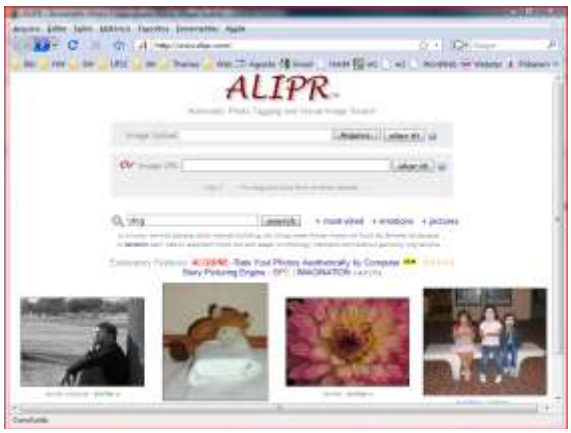
Prototype/project	MOE	Jasper[1]	TAP	Hybrid spreading activation	ISM	Librarian agent
By	Haffer and Reuther (2000)	Geisler et al. (2001)	Dubo et al. (2003)	Rocha et al. (2004)	Burton-Jones et al. (2003)	Stojanovic (2003)
Focus	WWW	WWW	WWW	WWW	WWW	WWW
Architecture	Stand-alone	Mets	Mets	Stand-alone	Mets	Stand-alone
Coupling	Tight	–	Loose	Tight	Loose	Tight
Transparency	Interactive	Transparent	Hybrid	Transparent	Hybrid	Interactive
User context	None	Hard-coded	Hard-coded and learning	None	None	Learning
Query modification	Manually	Conj. augm.	Ontology-part graph-based Document-part unclear	Graph-based	All sorts of query rewriting	Conj. learning Conj. augm. Substitution
Ontology structure	Hyponym Anonymous	–	Anonymous	Domain-specific	Hyponym Synonym Negation	Anonymous
Ontology technology	Proprietary	–	RDF	Unclear	ZebAI concepts + Word net	Proprietary

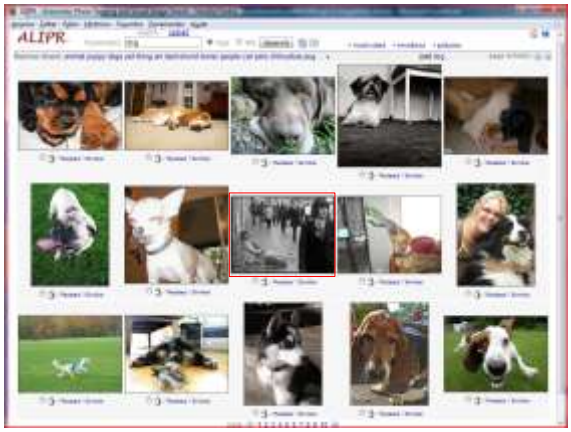
Comparação de abordagens (cont.)

Prototype/project	SCORE	TRUST	Audio data retrieval	Designer	Semantically poor effective
By	Sheth et al. (2002)	Janeiro et al. (2004)	Klein et al. (2004)	Mayfield et al. (2003)	–
Focus	Information system	Information system	IS for audio data retrieval	IS for image retrieval	–
Architecture	Stand-alone	Hybrid	Stand-alone	Stand-alone	Open
Coupling	Tight	Hybrid	Tight	Tight	Hybrid
Transparency	Interactive	Unclear	Transparent	Interactive	Hybrid
User context	Unclear	Hard-coded	None	Hard-coded	Hard-coded and learning
Query modification	Manually	Conj. augm.	Conj. augm. Substitution	Substitution	Open
Ontology structure	Hyponym Domain Specific	Hyponym Synonym Rest nucleus	Hyponym Synonym Meronym Instance_of	Hyponym Meronym Domain specific	Foundation-taxonomy
Ontology technology	Unclear	Proprietary	Proprietary	Proprietary	Gnutella[1]

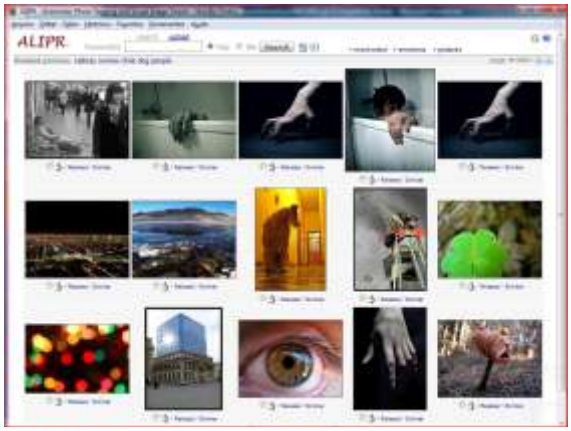
Buscas semânticas na Web hoje

- http://www.alipr.com (imagens)
- http://www.cognition.com
- http://www.deepdyve.com
- http://www.bing.com
- http://www.cuil.com
- http://www.freebase.com
- http://www.google.com
- http://www.kosmix.com
- http://www.hakia.com
- http://www.powerset.com
- http://www.sensebot.net
- http://developer.yahoo.com/searchmonkey
- http://swoogle.umbc.edu (ontologias)

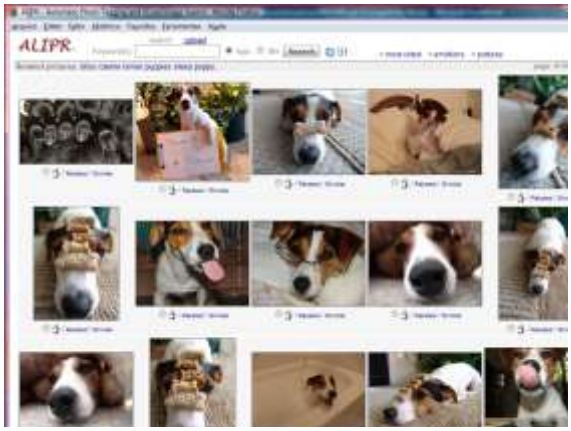



ALIPR


ALIPR

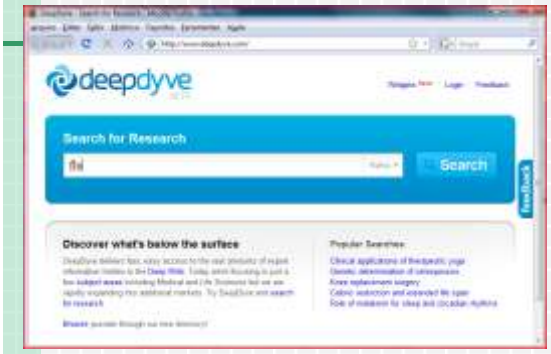

ALIPR


ALIPR


ALIPR


COGNITION
TRY IT FOR YOURSELF


MEDLINE COGNITION


deepdyve
Search for Research
Search
Discover what's below the surface


bing


bing


cuil
Sao Paulo
Pesquisa
SAO PAULO-SP
University of Sao Paulo
Sao Paulo Fashion
Sao Paulo Fashion Week
Sao Paulo State
Sao Paulo Hotel
Sao Paulo Hotels
Sao Paulo Brasilien
Sao Paulo Congonhas


A wealth of free data.


São Paulo Futebol Clube


Google
São Paulo


Google
lula


Google


Kosmix
sao paulo
Explore
Sao Paulo Airport
Sao Paulo Plane Crash
Sao Paulo Zoo
...
Top News
Top Buzz from Yahoo! Buzz


Kosmix
São Paulo
Explore
São Paulo
Video


hakia
sao paulo


hakia
sao paulo


Powerset
Ready. Powerset. Go.
Explore
Microsoft brings you a better way to search
bing

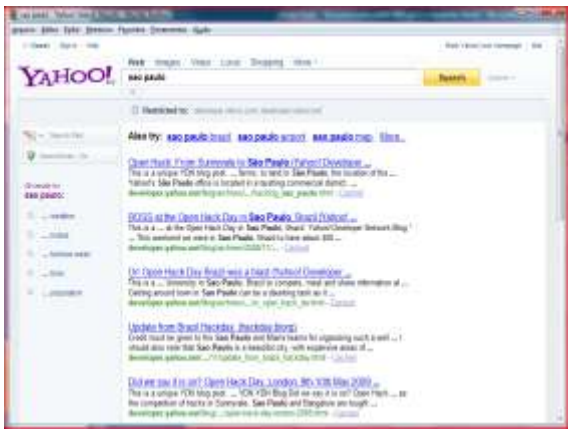

Powerset
Microsoft
São Paulo
bing


bing


SenseBot
The Search Engine that finds sense in a heap of Web pages
Search
Browser extensions
LiveSensor
Products | Services
NEW - Semantic API available!


SenseBot
The Search Engine that finds sense in a heap of Web pages
SAO PAULO


YAHOO! DEVELOPER NETWORK
SearchMonkey
Get Started
Read Documentation


YAHOO!
sao paulo

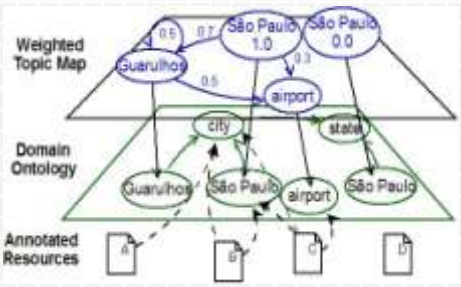

Swoogle
semantic web search
2007
Searching over 10,000 ontologies

Algumas questões em aberto

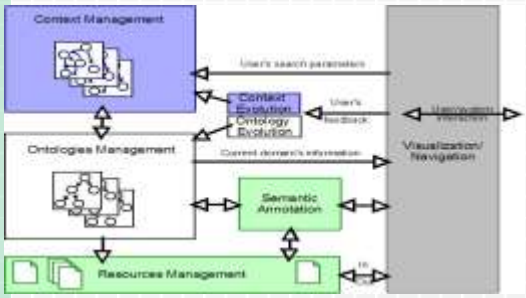
- Representação de contextos de usuários e mapeamento desses a ontologias
- Análise dos efeitos de modificações de consultas
- Meta-buscas semânticas
- Análise da aceitação dos usuários
- Adaptabilidade a diferentes ontologias
- Ranking dos resultados
- Integração com sistemas de gerenciamento de documentos/conteúdo
- Tratamento de dados multimída
- Interfaces homem-máquina
- Desempenho e escalabilidade

Contextual Semantic Search (CSS)

(D'Agostine, Fileto – SBBD2007)



General Architecture for CSS



The Weighted Topic Graph

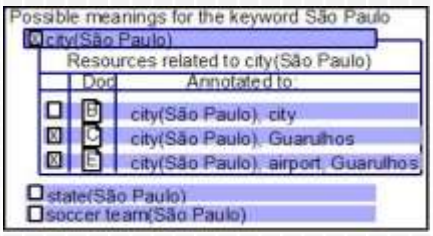
(D'Agostine, Fileto, Dantas, Gauthier - ICEIS 2008)

- A weighted topic graph is a graph ***TM(T, A)***
  - ***T*** is a set of topics (vertices)
  - ***A*** is a set of associations (edges)
- Let be an ontology ***O***
  - Each topic ***t*** in ***T*** corresponds to a ontology term ***o*** in ***O***
- Each topic ***t*** has a weight [0,1]. The sum from the weights from all topics with same name equals 1
- Each association ***a*** has a weight [0,1]. The sum from all the weights from all association departing a single topics equals 1

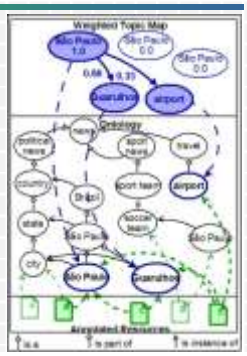
The Weights in the Topic Graph

- **Topic weights:** used to disambiguate
  - E.g., interest in *São Paulo* city instead of *São Paulo* state
- **Association weights:** used to semantically expand searches
  - E.g., Actual interest in Guarulhos and airport while searching for São Paulo city

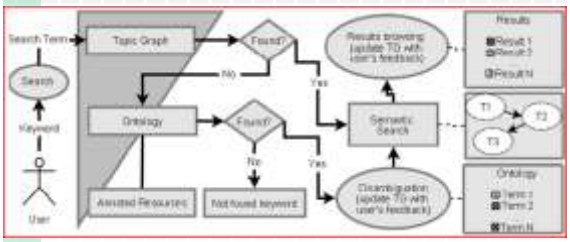
Capturing the User's Context

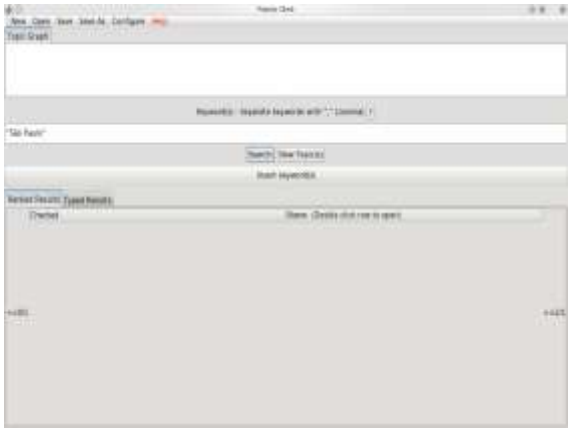


Generated Topic Graph with the Ontological User's Context

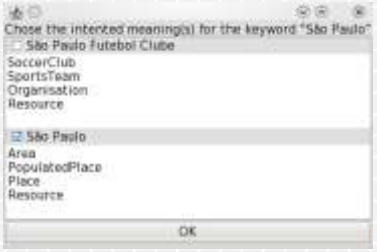


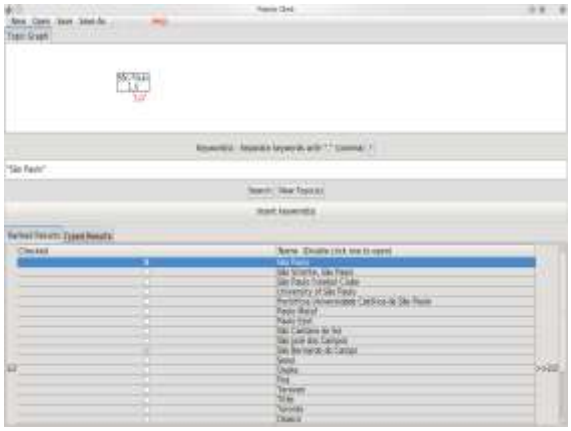
The Contextual Semantic Search Process

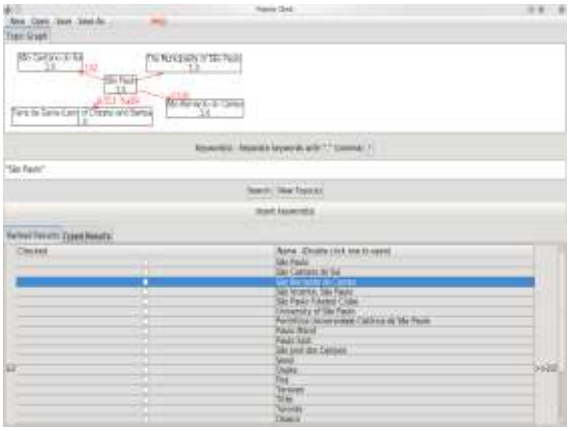




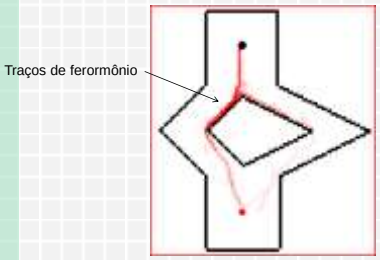
Desambiguação







Fundamento dos algoritmos de busca sobre contexto ontológico



Heurísticas de estigmergia (*Ant Colony Optimization - ACO*)

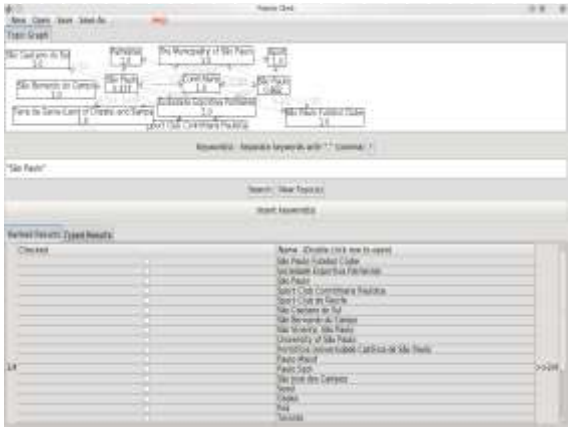
Algoritmo de busca semântica baseada em ACO

Entrada: palavras-chave[ ];

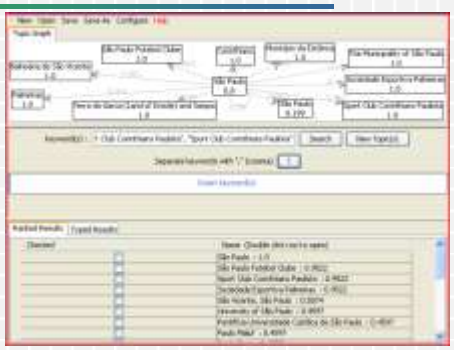
Dados: G(T,A), Onto, repositório,
limite_de_profundidade,
limite_mínimo_de_peso,
λ // fator de atenuação de peso;

início Algoritmo Busca Contextual 1 Estigmérica
resultados = busca_G_ou_Onto(*palavras-chave[]);*
atenuaG (*índice*) ; // λ ∈ *[0, 1]*
manutenção(coleta_feedback(*resultados));*

fim



Demo Praestro – Karina Fasolin



Alguns trabalhos relacionados

Name	Article	
Graupmann et al. 2005	The SphereSearch engine for unified ranked retrieval of heterogeneous XML and web documents	Searches through around given eleme to both contexts. Made tests with stress testing reg
Park e Cheyer 2006	Just For Me: Topic Maps and Ontologies	Manages knowledge in three layers: Topics, Knowledge Stru Documents. Not implemented (at least at the time of publication).
Michlmayr, et al. 2007	Adaptive User Profiles for Enterprise Information Access	Represents context as an individual graph. The graph is co based on tags used to semantically annotate the content which t considers to be relevant. However, the tags are only that, only labels. They are not tied-down to a formal definition.

benchmark for this kind of novel system.

Alguns trabalhos relacionados (cont.)

Name	Article	Comments	
Aleman-Meza et a. 2003	Context-Aware Semantic Association Ranking	Also defines context as regions. But the regions are static, defined in the ontology which is used by the system. So the contexts are defined the same for every single user.	
Mani and Sundaram 2007	Modeling user context with applications to media retrieval	Creates a graph representing the user's context. Each node corresponds to an instance in the ontology. The edges represent relations between the instances. Built to search trough multimedia documents. However, it does not associates the vocabulary the user uses with the instances.	
Challam et al. 2007	Contextual Search Using Ontology-Based User Profiles	Monitors the user's activity capturing content from open Internet Explorer, MS-Office and MSN. The captured information is stored and used to build a user's contextual profile bases on an ontology. The context is tightly coupled with the ontology.	NO IMAGES WERE PROVIDED
Vallet et al. 2006	Personalized Information Retrieval in Context	Stores context as a graph constructed based on the ontology used by the system. Similar to Mani and Sundaram, including its limitations (mapping the vocabulary), however if built for text documents.	
Sieg et al. 2007	Ontological User Profiles for Personalized Web Search	Similar to Challam et al., but it gets the context information from the ontology terms used to annotate the retrieved relevant context. Also limited by the structure of the ontology.	

Alguns trabalhos relacionados (cont.)

Name	Deals with objective knowledge (ontology)?	Deals with subjective knowledge (user's context)?	The knowledge management is transparent?	Idea / Implemented
Graupmann et al. 2005	X			IMPL.
Park e Cheyer 2006	X	X	X	IDEA
Michlmayr, et al. 2007		X	X	IMPL.
Aleman-Meza et a. 2003	X		X	IMPL.
Mani e Sundaram 2007	X	X	X	IMPL.
Challam et al. 2007	X	X?	X	IMPL.
Vallet et al. 2006	X	X?	X	IMPL.
Sieg et al. 2007	X	X?	X	IMPL. (incomplete)
Praestro	**X**	**X**	**X**	**IMPL. (prototype)**

Trabalhos futuros

- Processos parcialmente automatizados para geração e atualização de ontologias e anotações semânticas ???
- Experimentos para comprovar a eficácia das técnicas e ferramentas utilizadas
- Validação de soluções de problemas de busca semântica em diversos domínios de aplicação
- Determinação de níveis adequados de acoplamento entre contextos e ontologias e dessas com os recursos a serem recuperados

Notícias e análises

- http://news.cnet.com/new-search-engine-cuil-takes-aim-at-google/
- http://www.readwriteweb.com/archives/is_google_a_semantic_search_engine.php
- http://www.readwriteweb.com/archives/semantic_search_the_myth_and_reality.php
- http://www.pandia.com/sew/1262-top-5-semantic-search-engines.html
- http://mindset.research.yahoo.com/
- http://news.cnet.com/8301-13953_3-9982015-80.html
- http://www.searchenginejournal.com/askcom-focuses-on-semantic-search/8252/
- http://news.bbc.co.uk/2/hi/programmes/click_online/default.stm
- http://news.bbc.co.uk/2/hi/programmes/click_online/8144765.stm

Web semântica e buscas semânticas

- http://www.w3.org/2001/sw/
- http://semanticweb.org/
- http://en.wikipedia.org/wiki/Semantic_Web
- http://en.wikipedia.org/wiki/Semantic_search

Artigos básicos

- http://www.aifb.uni-karlsruhe.de/WBS/pha/publications/ontology-ir-ictir07.pdf
- http://portal.acm.org/citation.cfm?id=1359840
- http://espace.library.curtin.edu.au/R/?func=dbin-jump-full&object_id=116029&local_base=GEN01
- http://www.seco.tkk.fi/publications/2005/makela-semantic-search-2005.pdf

Referências em buscas semânticas

- Guha, R., McCool, R., and Miller, E. 2003. **Semantic search**. In *Proc. of the 12th international Conference on World Wide Web (WWW)*, Budapest, Hungary. ACM, New York, NY, 2003, 700-709.
- Reeve, L. and Han, H. 2005. **Survey of semantic annotation platforms**. In *Proceedings of the ACM Symposium on Applied Computing*, Santa Fe, New Mexico. ACM, New York, NY, 2005, 1634-1638.
- Eetu Makela. **Survey of Semantic Search Research**
- Mangold, C. **A Survey and classification of semantic search approaches.** Journal of Metadata Semantics and Ontology, 2(1), 2007.
- Stephan Bloehdorn, Philipp Cimiano, Alistair Duke, Peter Haase, Jörg Heizmann, Ian Thurlow, Johanna Völker. **Ontology-Based Question Answering for Digital Libraries**. ECDL 2007: 14-25.
- Hai Dong; Hussain, F.K.; Chang, E. **A survey in semantic search technologies**. In *2nd IEEE International Conference on Digital Ecosystems and Technologies.* 2008, 403-408

Algumas referências grupo UFSC

- D'Agostini, C. S. ; Fileto, R. **Capturing Users' Preferences and Intentions in a Semantic Search System**. In: *21st International Conference on Software Engineering & Knowledge Engineering (SEKE)*, Boston, 2009. p. 587-591.
- D'Agostini, C. S. ; Fileto, R. ; Dantas, M. A. R. ; Gauthier, F. A. O. **Contextual Semantic Search - Capturing, using the User's Context to Direct Semantic Search**. In: *10th International Conference on Enterprise Information Systems (ICEIS)*, Barcelona, Spain, 2008. v. SAIC. p. 154-159.
- D'Agostini, C. S. ; Fileto, R. **Capturing and managing user context for improving information retrieval.** In: *Workshop de Teses e Dissertações do Simpósio Brasileiro de Bancos de Dados (WTDBD/SBBD),* Campinas: Unicamp, Brazil, 2008.
- Vian, J. ; Silveira, R. A. ; Fileto, R . **Proposal of a Multi-agent System for Indexing and Recovery applied to Learning Objects**. In: *9th IFIP World Conference on Computers in Education (WCCE)*, Bento Goncalves, Brazil, 2009.

Perguntas?