Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents

Cláudio Gottschalg Duque[1]

(klauss@eci.ufmg.br)

Marcello Peixoto Bax

(bax@eci.ufmg.br)

ESCOLA DE CIÊNCIA DA INFORMAÇÃO - ECI (www.eci.ufmg.br)
UNIVERSIDADE FEDERAL DE MINAS GERAIS – UFMG (www.ufmg.br)

1- Supported by CNPq (www.cnpq.br)

Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents

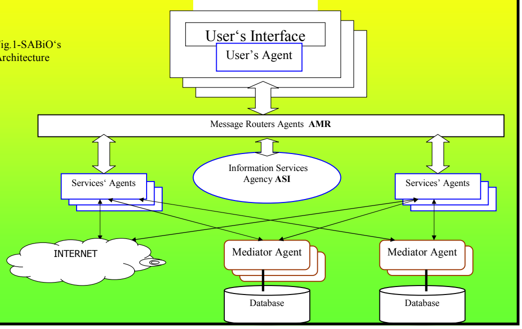
- Cognitive Sciences can help to develop a robust Information Retrieval System that uses linguistic and knowledge information extracted from texts that are indexing, allowing more efficiency in answer user's queries.
- We present an index proposal, as part of SABiO's Project, whose approach involves the application of specific linguistic theories and the use of Protégé to develop an ontology based in the terms extracted from texts.

Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents

- The SABiO Project aims to put in operation an experimental Digital Library that will serve his users with Information Retrieval Agents and Interface Agents.
- The goal is develop new concepts of Distributed Information Architectures, using straight and indirect collaboration tools between users and Digital Libraries.
- These tools will help Digital Libraries to answer the challenges and opportunities created by the recent technologies of Information and Communication.

Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents

Fig.1-SABiO's Architecture



Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents

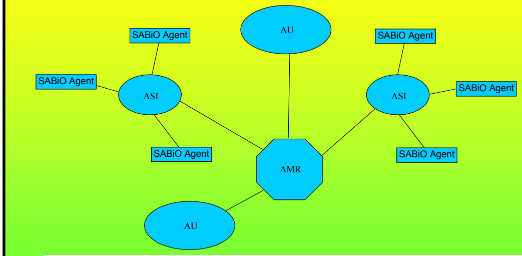


Fig.2- SABiO's Network

ASI – Information Services Agency AMR – Message Routers Agent AU – User's Agent

Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents

Information Retrieval System in SABiO Project:

– Natural Language Processing System (NLPS).

– Ontology Processing System (OPS).

– Index Processing System (IPS).

Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents

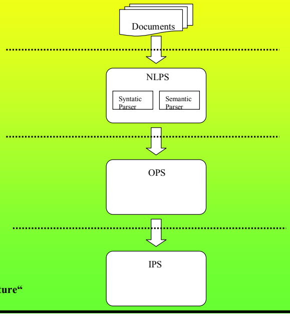


Fig.3 IRS "Big Picture"

Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents

Natural Language Processing System (NLPS).

- The NLPS would be robust:
  - capable to make effective positive difference in relation statistical approaches used in traditional IRS.
- NLPS process documents, extracting descriptions using:
  - Syntactic Module System (SYMS).
  - Semantic Module System (SEMS).

Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents

Natural Language Processing System (NLPS)

- SYMS and SEMS

- These modules are used for extract some possible concepts units that will be put in the OPS.

- The Syntactic Module System (SYMS) is responsible for the content words (verbs, nouns, etc) and the functional words (prepositions, conjunctions, etc) identification.

- The Semantic Module System (SEMS) identifies relevant semantic issues, primitive concepts.

Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents

SYMS

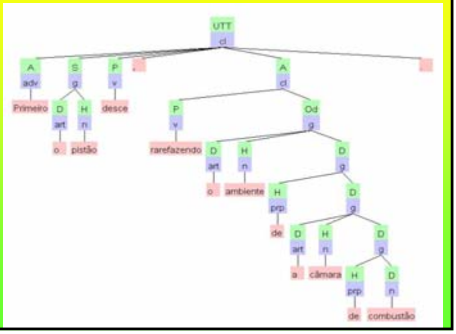
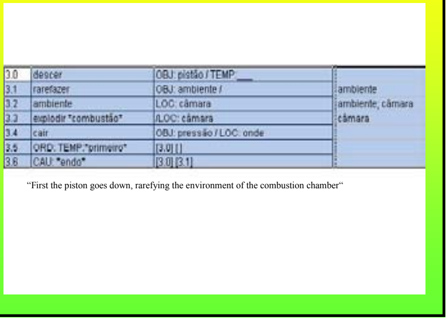


Fig.4- "Primeiro o pistão desce, rarefazendo o ambiente da câmara de combustão (http://visl.hum.sdu.dk/visl/) First the piston goes down, rarefying the environment of the combustion chamber"

Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents



3.0	descer	OBJ: pistão / TEMP: ___	
3.1	rarefazer	OBJ: ambiente /	ambiente
3.2	ambiente	LOC: câmara	ambiente, câmara
3.3	explodir "combustão"	/LOC: câmara	câmara
3.4	cair	OBJ: pressão / LOC: onde	
3.5	ORD: TEMP:"primeiro"	[3.0] []	
3.6	CAU: "endo"	[3.0] [3.1]	

"First the piston goes down, rarefying the environment of the combustion chamber"

Fig.5 – Semantic Parser

Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents

Use of Concepts and Ontologies to Index

- We use ontologies because they are structured in a way that goes considerably beyond the possibilities offered by others classification systems like thesauri.

- Syntactic contexts and Semantic roles are used to infer that some terms are really concepts and can be extracted from texts.

Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents

Use of Concepts and Ontologies to Index

- The idea is that structured index concepts generated by means of extensive analysis result in a better performance for answer user's queries.
- The goal is, for one specific text collection, create and update the ontology automatically.

Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents

Use of Concepts and Ontologies to Index

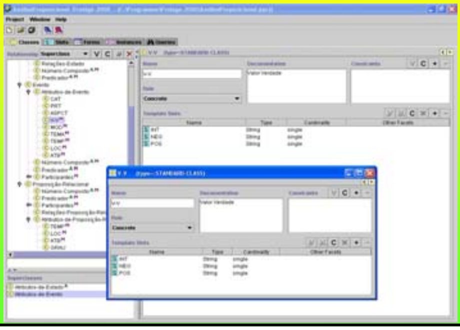
- Based on text terms and possible relationships between terms, we create the ontology in Information Science domain.
- The organization of the System is modular, so the natural-language part (NLPS) is strictly separated from the Index Processing System (IPS) (i.e., the module that has access to the ontology).

Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents

Use of Concepts and Ontologies to Index

- It means that all linguistic process that are necessary to create the ontology are processed in separate.
- Theoretic, the available knowledge is enough to produce a clear and unambiguous analysis, if not, the NLPS should return blocks of possible options that can be used by the OPS and the IPS.

Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents



V-V (type=STANDARD-CLASS)

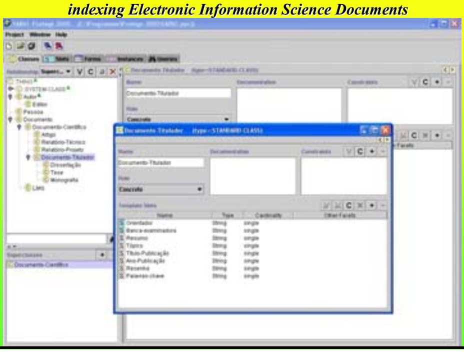
Name: V-V

Documentation: Valor Verdade

Role: Concrete

Name	Type	Cardinality	Other Facets
AFT	String	single	
NEG	String	single	
POS	String	single	

Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents



Documento-Titulador (type=STANDARD-CLASS)

Name: Documento-Titulador

Role: Concrete

Name	Type	Cardinality	Other Facets
Orientador	String	single	
Banca-examinadora	String	single	
Resumo	String	single	
Tópico	String	single	
Título-Publicação	String	single	
Ano-Publicação	String	single	
Resenha	String	single	
Palavras-chave	String	single	

Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents

Discussion

- We believe that we can propose contributions for syntactic parsers and semantic parsers in Information Retrieval applications.
- Syntactic context notation (aiming to improve the identification of semantically related words), combined with the ability of ontology tools (to construct and visualize multiple relationships), permits the elimination of large and complex ontology making the IRS more fast and efficient.

Use of Protégé in SABiO, a Brazilian Information Retrieval System for indexing Electronic Information Science Documents

Production

- 7 published articles.
- 1 Master Thesis and 2 in progress (publication in 2003).
- 1 Doctoral Dissertation in progress.
- 6 Technical Report (related orientations of scholarship holders under graduation students).
- 8 Under Graduation monographs.

(http://cuba.eci.ufmg.br/Bax/Projetos/Sabio/default.htm)