Universidade Católica de Pelotas

Stanley Loh (líder)

Evento PLN
Janeiro de 2008

Grupo de Pesquisa

GPSI

G rupo de

P esquisa em

S istemas de

I nformação

Pesquisadores

- Professores
  - Stanley Loh
  - Daniel Lichtnow
  - Ramiro Saldaña

Financiamento

- CNPQ, edital PDPG-TI e outros
- CNPQ, edital CT-INFO, junto com UFRGS
- FAPERGS, edital PRONEX, junto com UFRGS

Parceiros e Colaboradores Externos

- Prof. Dr. José Palazzo M. de Oliveira – UFRGS
- Prof. Dr. Leandro Krug Wives – UFRGS
- Profa. Dra. Renata Vieira – UNISINOS
- Profa. MSc. Fabiana Lorenzi - ULBRA
- Profa. Dra. Viviane Orengo – UFRGS
- Prof. Dr. Francesco Ricci – ITC/Itália

Projetos Principais

- SisRecCol:
  - Sistema de Recomendação para Apoio à Colaboração
  - Finalizado em 2005
- SisRec 2:
  - Sistema de Recomendação para Apoio à Aprendizagem e Busca de Informações
  - Finalizado em 2007
- Tag CloudYAH
  - Hierarquia de Tag Clouds
  - Início em 2008

Tema Atual

- Sistemas de Recomendação
  - Oferecer produtos ou informações personalizadas, de acordo com perfil do usuário

- Aplicações
  - Marketing e Comércio Eletrônico
  - Busca na Web
  - Bibliotecas Digitais
  - Apoio à aprendizagem e ensino

Foco

- Recomendação a partir da análise de textos
  - Documentos eletrônicos
    - artigos científicos, páginas web
  - Currículos vitae
    - Formato Lattes, XML
  - Descrições textuais de produtos

Temas Relacionados

- Classificação de textos
- Text Mining
  - Análise qualitativa e quantitativa de coleções de textos
- Ontologias
- Busca na Web

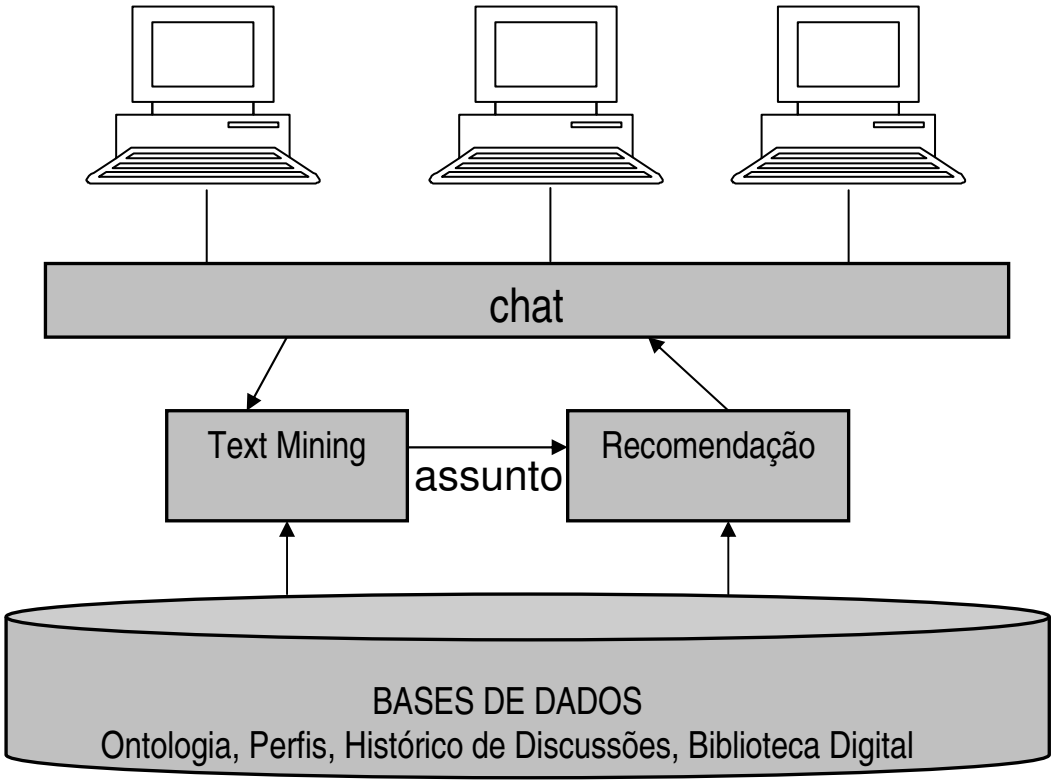
Tipo de Abordagem

- Análise Probabilística de Textos

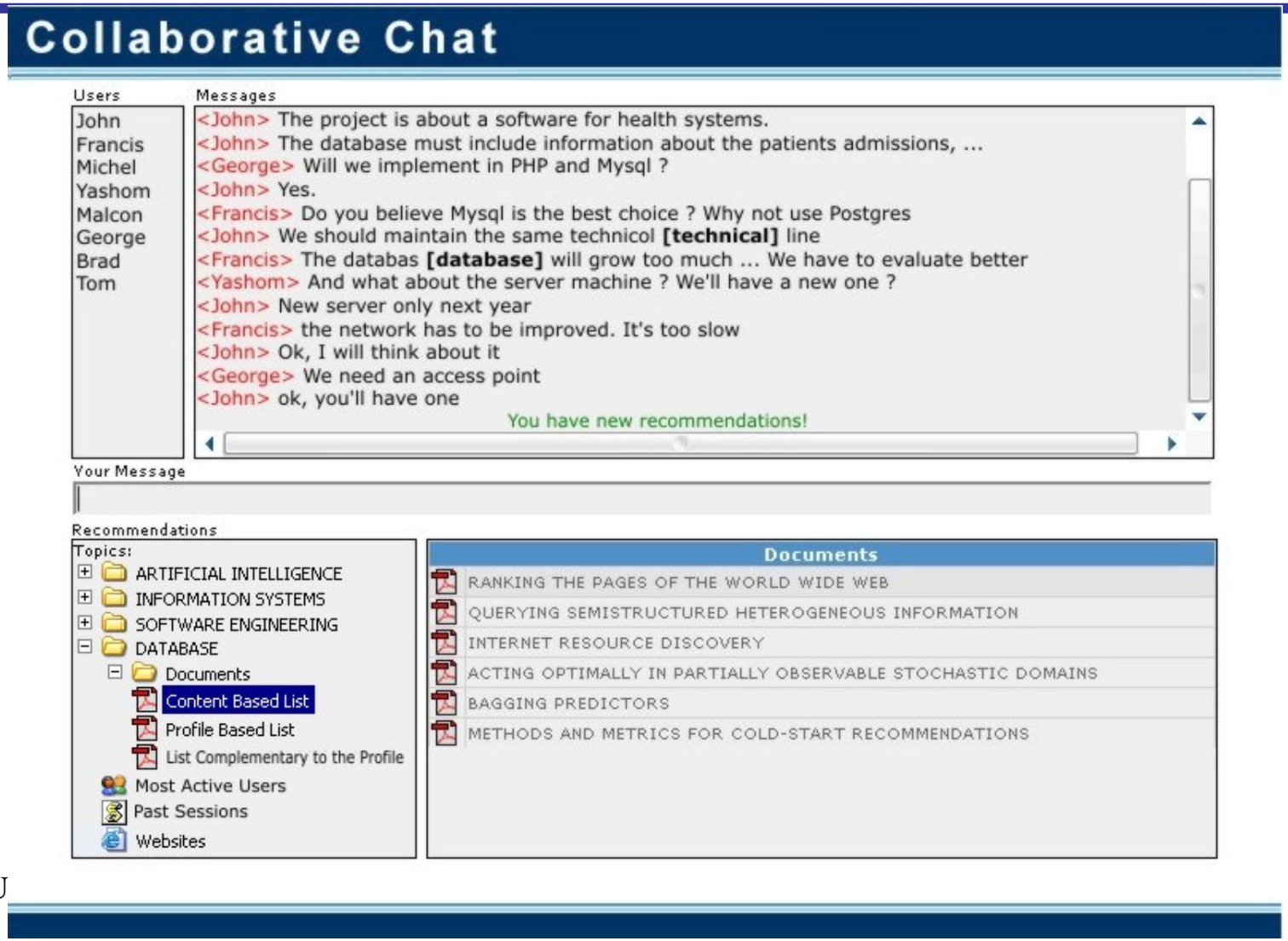
- Procuramos evitar:
  - Trabalho humano
  - Regras muito específicas para um domínio
  - Utilização de análise sintática (diminuir esforço de análise dos textos)

Projeto SisRecCol

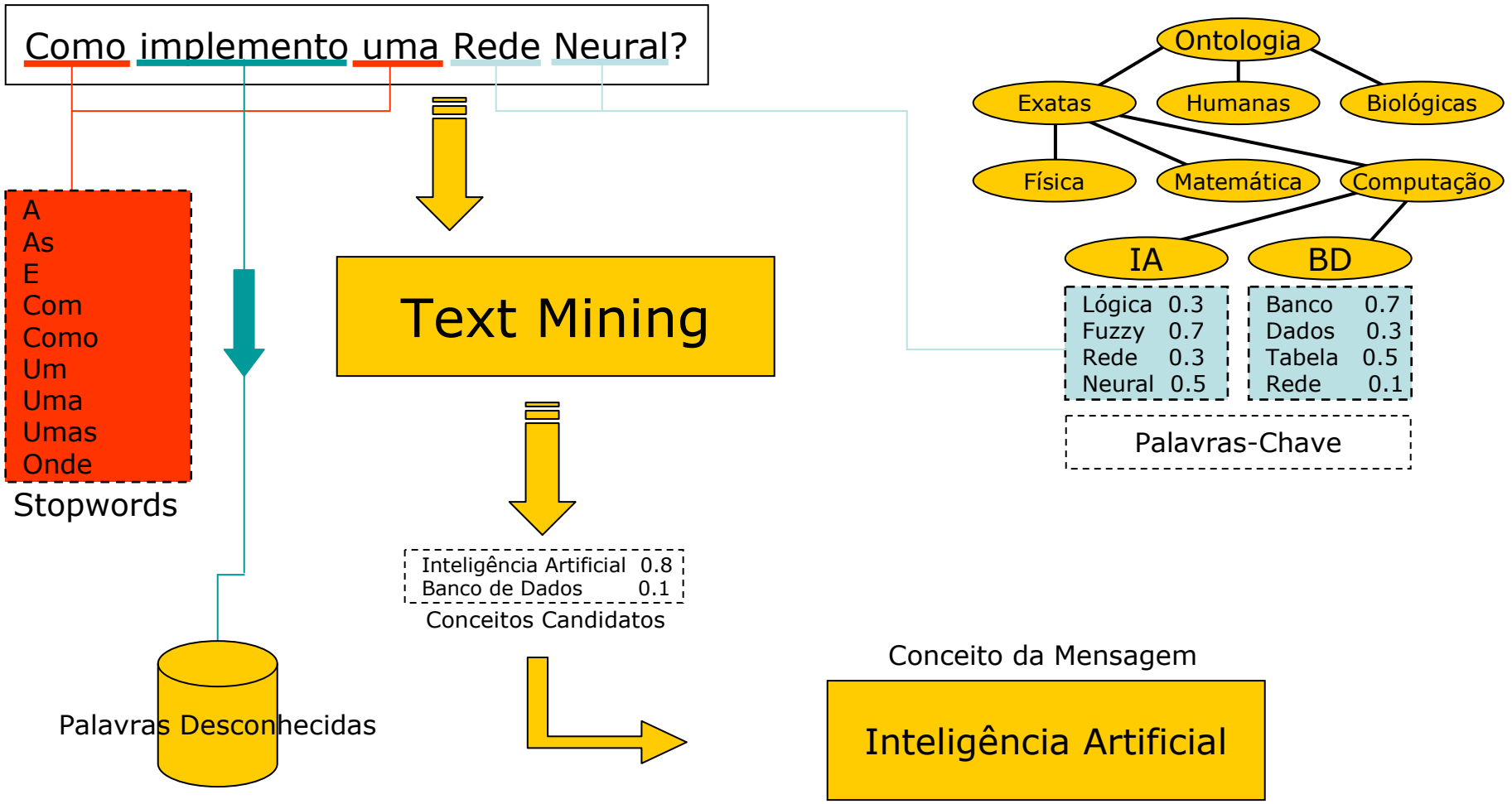
- Arquitetura básica



Tela Principal



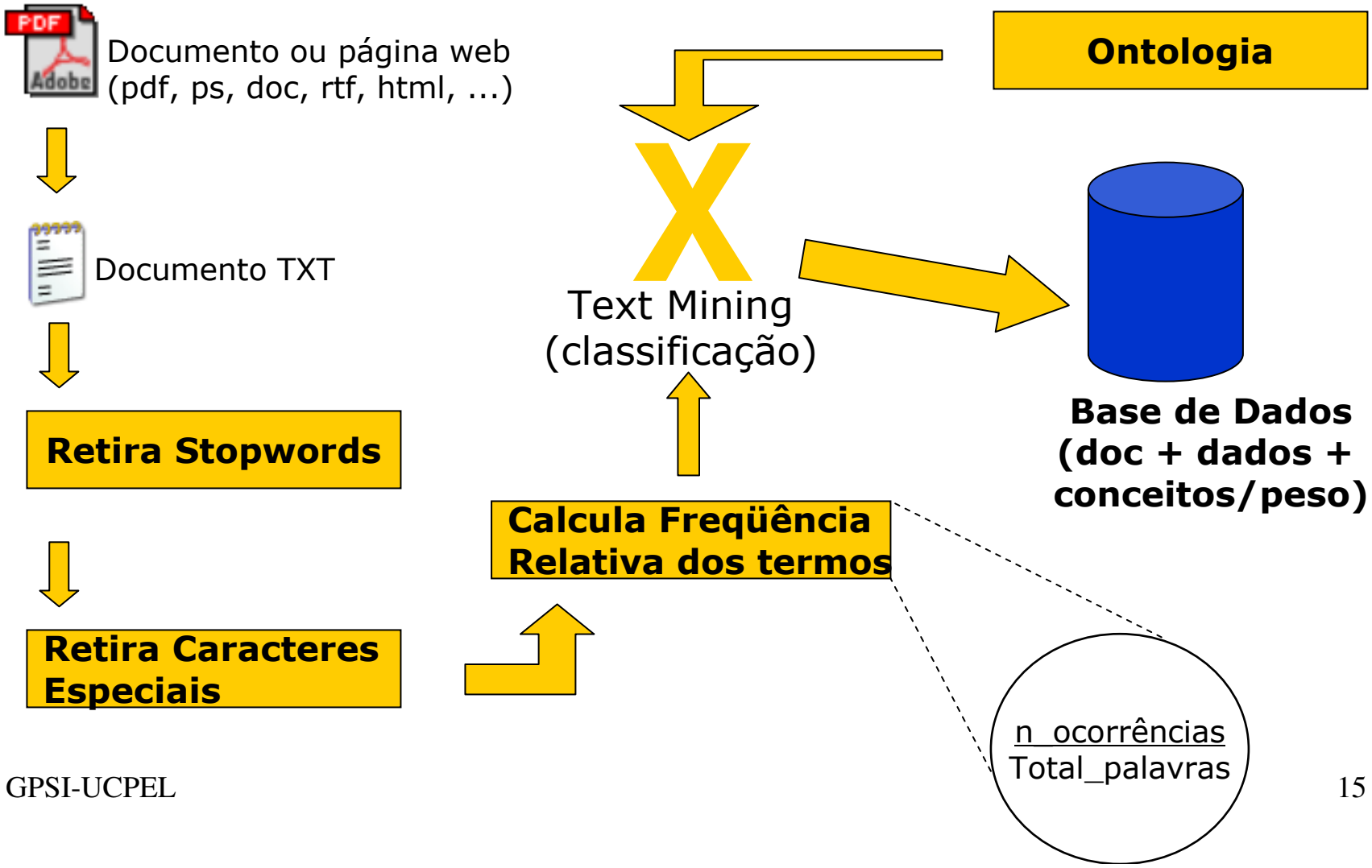
Módulo de Text Mining



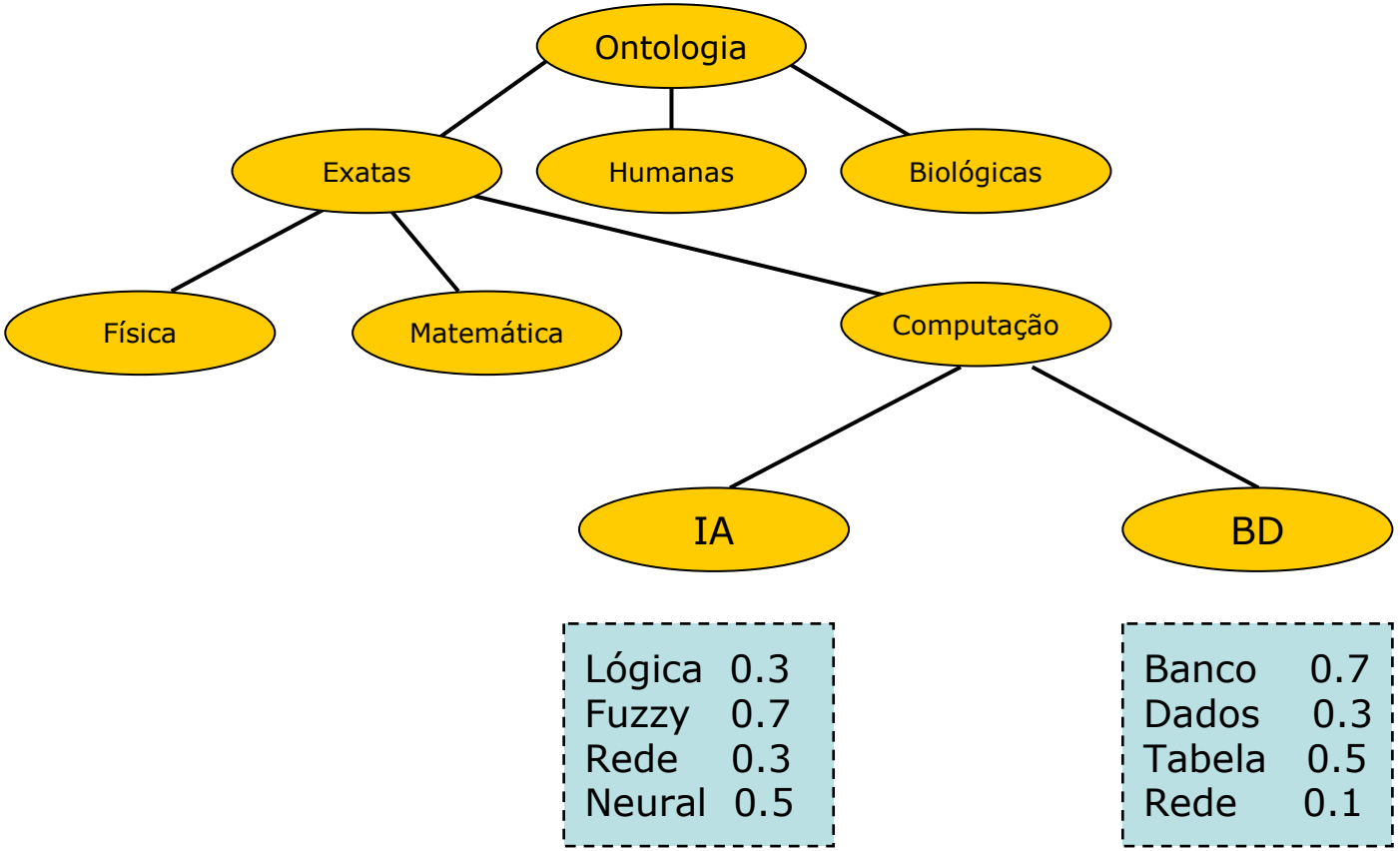
Identificação de Assuntos

- Método probabilístico
- Uso de limiar para cortar assuntos (definido por testes)
- Cada mensagem ➔ 1 assunto identificado (= conceito da ontologia)
- Grupo de mensagens ➔ assunto corrente da discussão (obs: evita ambigüidades)
- Uma discussão (sessão do chat) pode conter vários assuntos

Biblioteca Digital

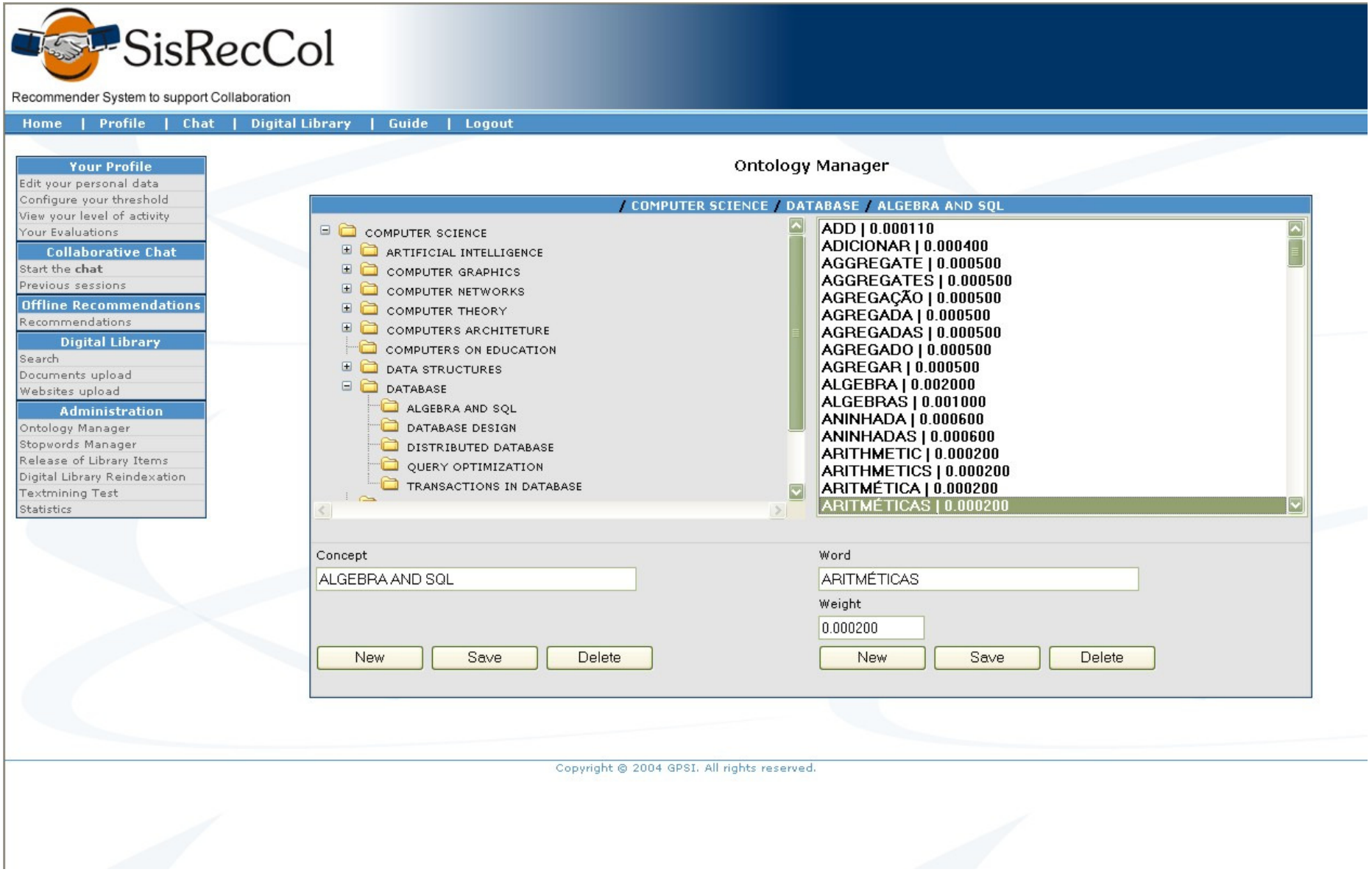


Ontologia de Domínio



Criação da Ontologia de Domínio

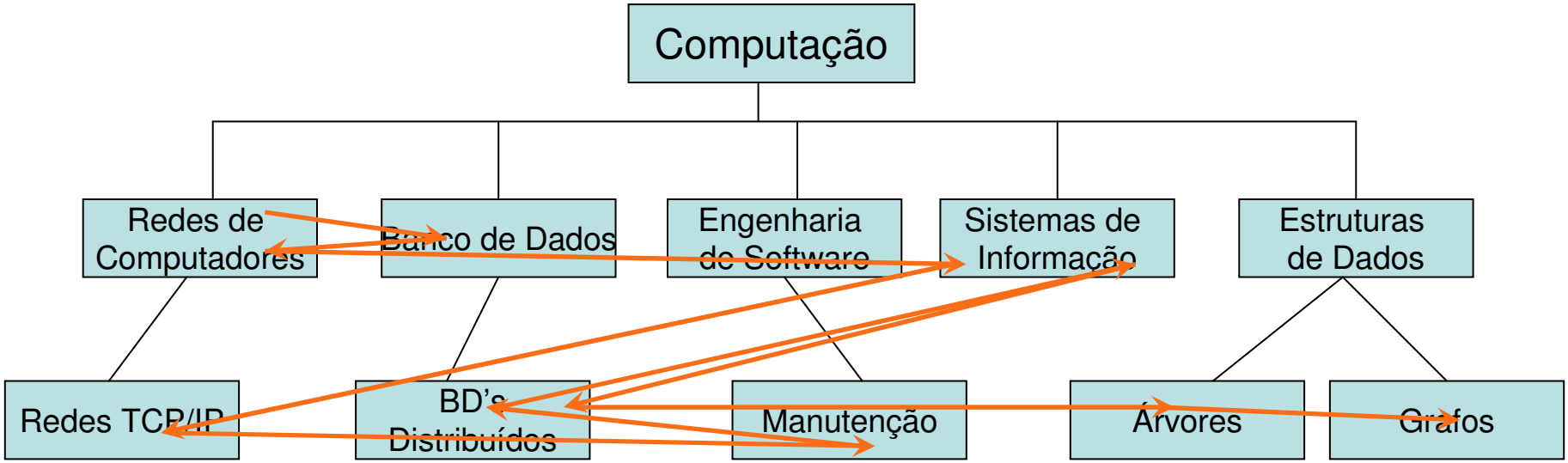
- Participação de especialistas no domínio
- Ferramentas automatizadas para inclusão de conceitos, termos, hierarquia, etc.
- Aprendizado supervisionado
  - especialista seleciona textos sobre um conceito
  - software identifica termos e pesos para conceitos
- Clustering de documentos e textos
  - identificar automaticamente conceitos, relações e termos
- Trabalhos de Thyago Borges e Gustavo Piltcher


SisRecCol
Recommender System to support Collaboration
Home | Profile | Chat | Digital Library | Guide | Logout
Your Profile
Edit your personal data
Configure your threshold
View your level of activity
Your Evaluations
Collaborative Chat
Start the chat
Previous sessions
Offline Recommendations
Recommendations
Digital Library
Search
Documents upload
Websites upload
Administration
Ontology Manager
Stopwords Manager
Release of Library Items
Digital Library Reindexation
Textmining Test
Statistics
Ontology Manager
/ COMPUTER SCIENCE / DATABASE / ALGEBRA AND SQL
COMPUTER SCIENCE
ARTIFICIAL INTELLIGENCE
COMPUTER GRAPHICS
COMPUTER NETWORKS
COMPUTER THEORY
COMPUTERS ARCHITETURE
COMPUTERS ON EDUCATION
DATA STRUCTURES
DATABASE
ALGEBRA AND SQL
DATABASE DESIGN
DISTRIBUTED DATABASE
QUERY OPTIMIZATION
TRANSACTIONS IN DATABASE
ADD | 0.000110
ADICIONAR | 0.000400
AGGREGATE | 0.000500
AGGREGATES | 0.000500
AGREGAÇÃO | 0.000500
AGREGADA | 0.000500
AGREGADAS | 0.000500
AGREGADO | 0.000500
AGREGAR | 0.000500
ALGEBRA | 0.002000
ALGEBRAS | 0.001000
ANINHADA | 0.000600
ANINHADAS | 0.000600
ARITHMETIC | 0.000200
ARITHMETICS | 0.000200
ARITMÉTICA | 0.000200
ARITMÉTICAS | 0.000200
Concept
ALGEBRA AND SQL
New
Save
Delete
Word
ARITMÉTICAS
Weight
0.000200
New
Save
Delete
Copyright © 2004 GPSI. All rights reserved.

Análise das Discussões

- Tema principal da discussão e temas periféricos
- Participação dos usuários (grau de atividade)
  - quem mais envia mensagens
  - assuntos de interesse
- Mineração utilizando método de análise temporal

Rota da Discussão

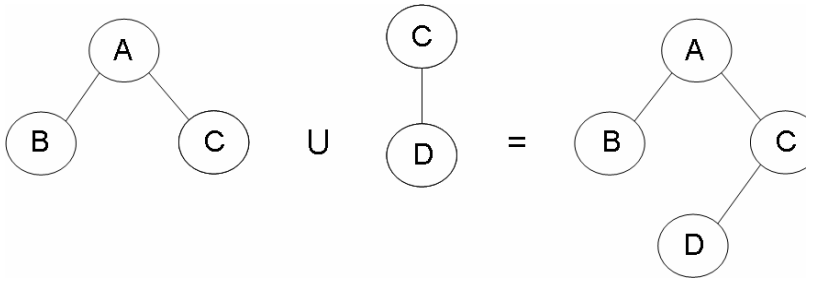


Grafos Representativos de Textos

- Cada texto ➔ 1 grafo
- Nodos = palavras
- Arestas = relações de proximidade entre as palavras
- Grau ou peso nas arestas
  - Fórmula
  - Análise da proximidade

Grafos Representativos de Textos

- Comparações entre grafos (textos)
  - união (todo o conhecimento de uma pessoa ou biblioteca)
  - intersecção (o que tem em comum)
  - diferença (o que eu não sei, que a biblioteca sabe)
  - busca por subgrafo



Grafos Representativos de Textos

- Outro objetivo principal:
  - Encontrar relações entre palavras

- Aplicações
  - Expansão semântica de buscas
  - Encontrar palavras sinônimas
  - Identificação do contexto de documentos
    - Ex: Google sugere termos (correção ortográfica nas buscas) pela análise dos termos relacionados

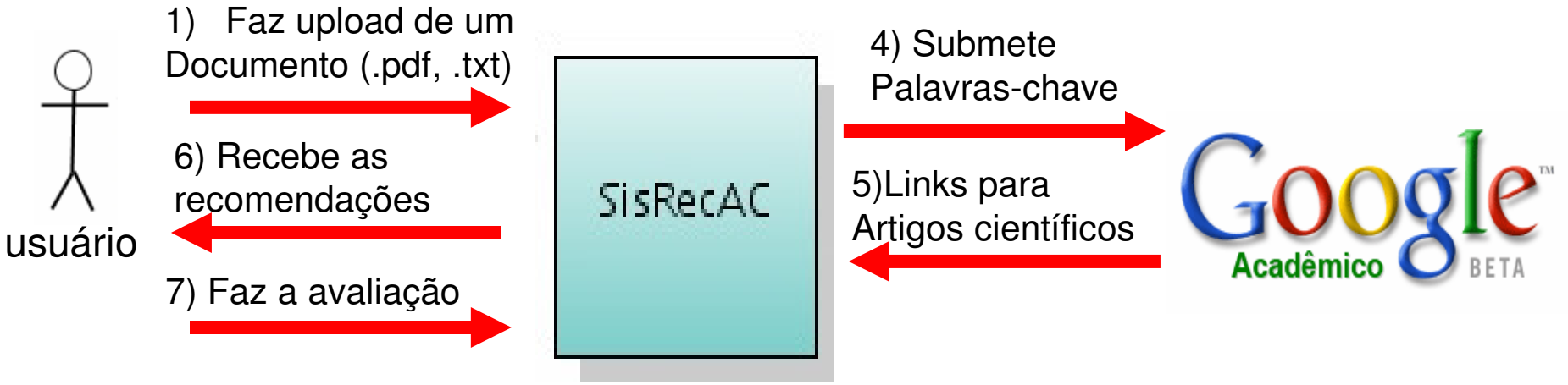
SisRecAC

- Sistema de recomendação de artigos científicos (SisRecAC)
- Baseado no paradigma de "*query by example*"
- É um sistema de metabusca

SisRecAC

	Consultas Tradicionais	**Query by example**	**Recomendação**
Descrição	O usuário deve saber informar corretamente as palavras-chave	O usuário informa um exemplo do que precisa	O sistema constrói um perfil dos usuários (filtragem colaborativa, baseado em conteúdo)
Exemplos de Sistemas	Google, Yahoo, outros	SisRecAC	Movielens, Grouplens, diversos sistemas de e-commerce
Problemas	ASK (Estado Anômalo de Conhecimento)	Ter o documento de exemplo	Partida a frio (Cold start), pouca possibilidade de surpresa (serendipity)

SisRecAC



2) Escolhe um método

3) Extrai palavras-chave do Documento


UNIVERSIDADE CATÓLICA DE PELOTAS
GPSI
RESEARCH GROUP ON INFORMATION SISTEMS
CATHOLIC UNIVERSITY OF PELOTAS
Home
Documentos
Estatística
Sair
competitiva povo search_engine j2me php tagging framework orientação_objetos educação blog retrieval search Scientific
sistema_recomendação Information recomendação rails
framework utilizada 4 vezes. Relações: php java web programming javascript
social_network ruby book keyphrase sisrecac comunicação ontology web web2.0 redes aulanet
biblioteca pedagogia folksonomy virtuais ajax keyphrase_extraction conhecimento engine sumarização extraction
resumos multiagentes google Papers java inteligência textos coletiva folksonomia metodologia informação recommending
sistemas aprendizagem gestão ensino wiki automática
Visualizar Documentos
Acesso a Bancos de Dados através de Celulares
28/09/2007 (Download) Tags: banco de dados java celular
Excluir documento
Recomendar para amigo
Upload
Folcsonomia: Vocabulário Descontrolado,Anarquitetura da Informação ou Samba do Crioulo Doido?
27/09/2007 (Download) Tags: folcsonomia folksonomia folksonomy
Excluir documento
Recomendar para amigo
Tags
A COMUNICAÇÃO EDUCATIVA EM AMBIENTES VIRTUAIS
16/09/2007 (Download) Tags: ambientes virtuais comunicação educação
Excluir documento
Recomendar para amigo
ANALISE DE PROJ.DE ALGORITMOS
16/09/2007 (Download) Tags: algoritmos


Home
Documentos
Estatística
Sair
Visualizar Recomendações
Upload
Voltar
Folcsonomia: Vocabulário Descontrolado,Anarquitetura da Informação ou Samba do Crioulo Doido?
27/09/2007 (Download) Tags: folcsonomia folksonomia folksonomy
Comentário:
Formação de Grupos de Trabalho Utilizando Agentes de Software
Recomendação 1
trabalhadores
deve ser fornecido tanto para a criação dos grupos de trabalho assim ...
LM Cunha - 2002 - groupware.les.inf.puc-rio.br
Este documento é relevante no contexto do documento de origem ?
Totalmente relevante.
Parcialmente relevante.
Irrelevante.
Avaliar
Recomendação 2
...EM LINGÜÍSTICA APLICADA E ESTUDOS DA LINGUAGEM TÓPICO EM DESCRIÇÃO DE LÍNGUAS: LINGÜÍSTICA DE ...
neutra, positiva ou negativa, podendo ser revelada por meio ... ser considerados na
construção e na descrição de um ... que leva o aluno a participar ativamente da ...
O ESCOLHIDA - lael.pucsp.br
Este documento é relevante no contexto do documento de origem ?
Totalmente relevante.
Parcialmente relevante.
Irrelevante.
Avaliar
Recomendação 3
A Adaptação do Ambiente AulaNet para dar Suporte a Grupos de Aprendizagem e
groupware como o software multi-usuário que apóia ... posteriormente faz-se uma breve
descrição da especificação ... eo seu relacionamento com grupos de trabalho ...

SisRecAC

- **Métodos de extração de palavras-chave**
  - Palavras mais freqüentes no texto (termos simples)
  - Expressões mais freqüentes (por análise estatística)
  - Palavras do título
  - "*tags*" informadas pelos usuários
  - Expansão de tags usando relações entre tags conforme sites populares

Tags com Expansão Semântica

- Alguns sites informam as tags mais utilizadas em conjunto

TAGS RELACIONADAS COM framework

bibsonomy:
java javascript software web ajax programming development web2.0 jQuery cms python webdesign develop opensource php collaboration t
testing learning design imported computing java_ee collaborative plugin architecture layout library .net webdev Baum resources unit analys
j2ee science eclipse visualization css apache security microsoft cognition ruby rdf graph c community xml BibSonomy is offere

technorati:
php ajax java programming web javascript opensource development software .net

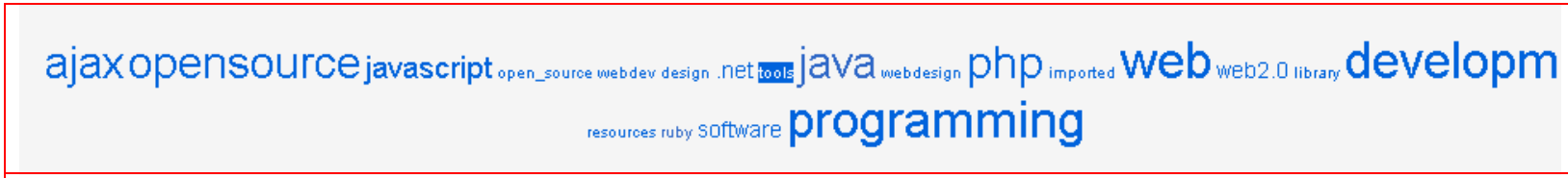
del.icio.us:
framework programming java ajax php web javascript .net development opensource

stumbleupon:
ajax development open source php programming web

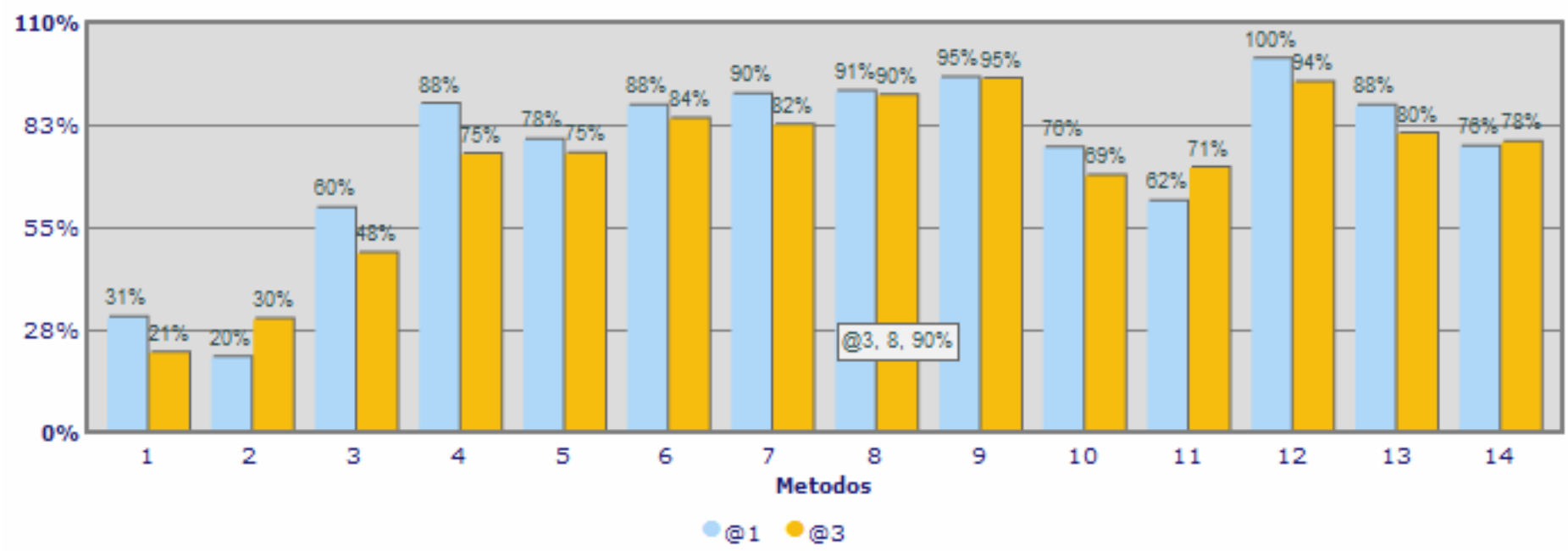
squidoo:
framework programming java web frameworks php opensource development architectural taxonomy

bluedot:
ajax development java Javascript library mvc open_source opensource php programming Python rails reference ruby tools tutorial web we
webdev

netvouz:
framework ajax development javascript open_source programming webdesign code resources web design web2.0 oop java project spf co
rad apps inspiration google api networking mail jsf lists digg dev scripts imported howto openstandard applications tutorials opensource ipta
blog iphone cool gadgets programmering tools software resource gears 2.0 study netfilter tips



Resultados



1 a 3: expressões mais freqüentes

4 a 9: termos simples mais freqüentes

10: palavras do título

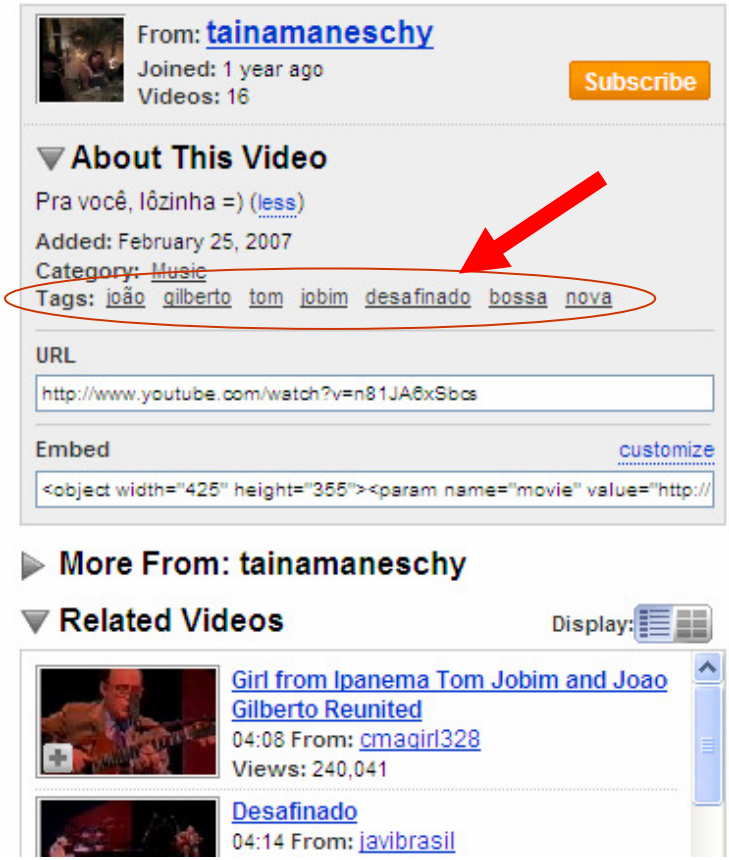
11: tags informadas pelo usuário

12 a 14: tags expandidas (1, 2 ou 3 tags a mais)

Tagsonomias ou Folksonomias

Folk + Taxonomia = Folksonomia

João Gilberto & Tom Jobim - Desafinado



Tagsonomias ou Folksonomias

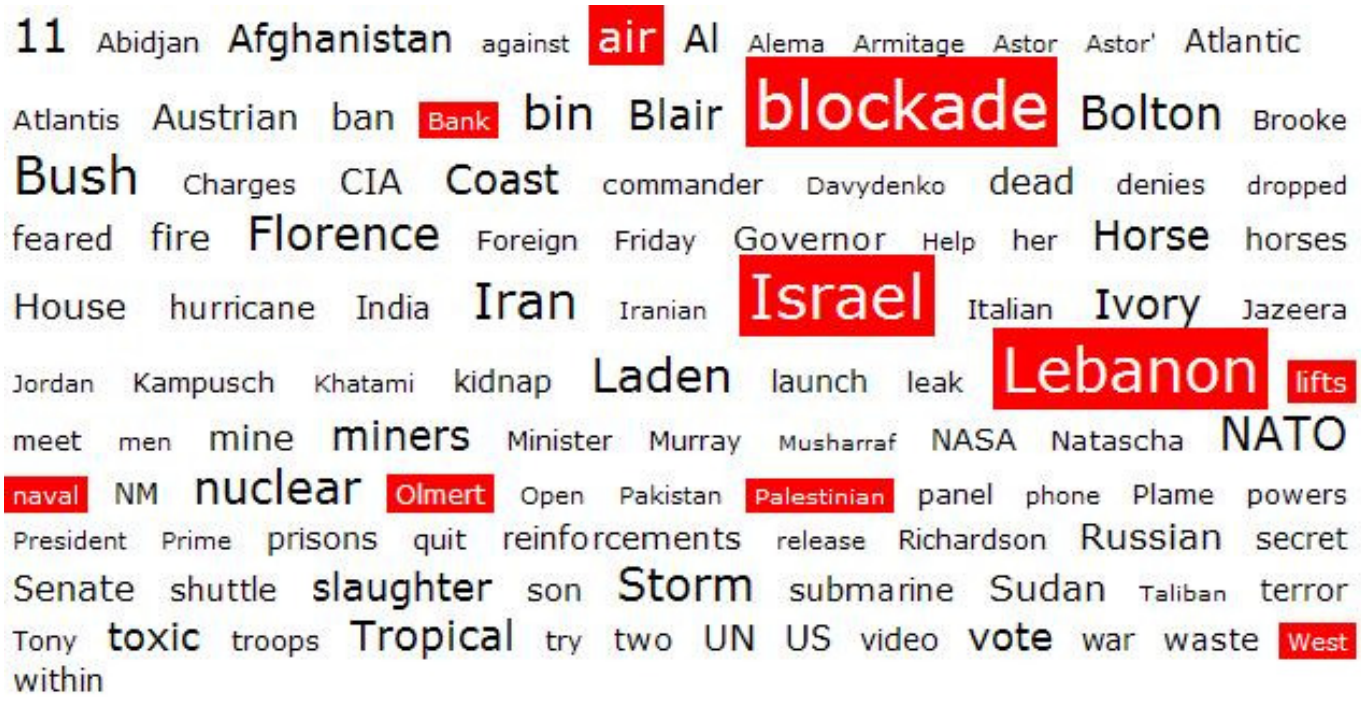
Folksonomias	Ontologias e Taxonomias
Feitas por leigos Maior flexibilidade Novas classes, inclusive pela combinação de outras Permite identificar assuntos principais e periféricos Problemas com polisemia e sinônimos	Feitas por especialistas Estrutura mais rígida Uso somente de classes já existentes Às vezes, pertinência a somente uma classe Confusão entre assunto principal e periféricos Resolve melhor problema de sinônimos

Tagsonomias ou Folksonomias

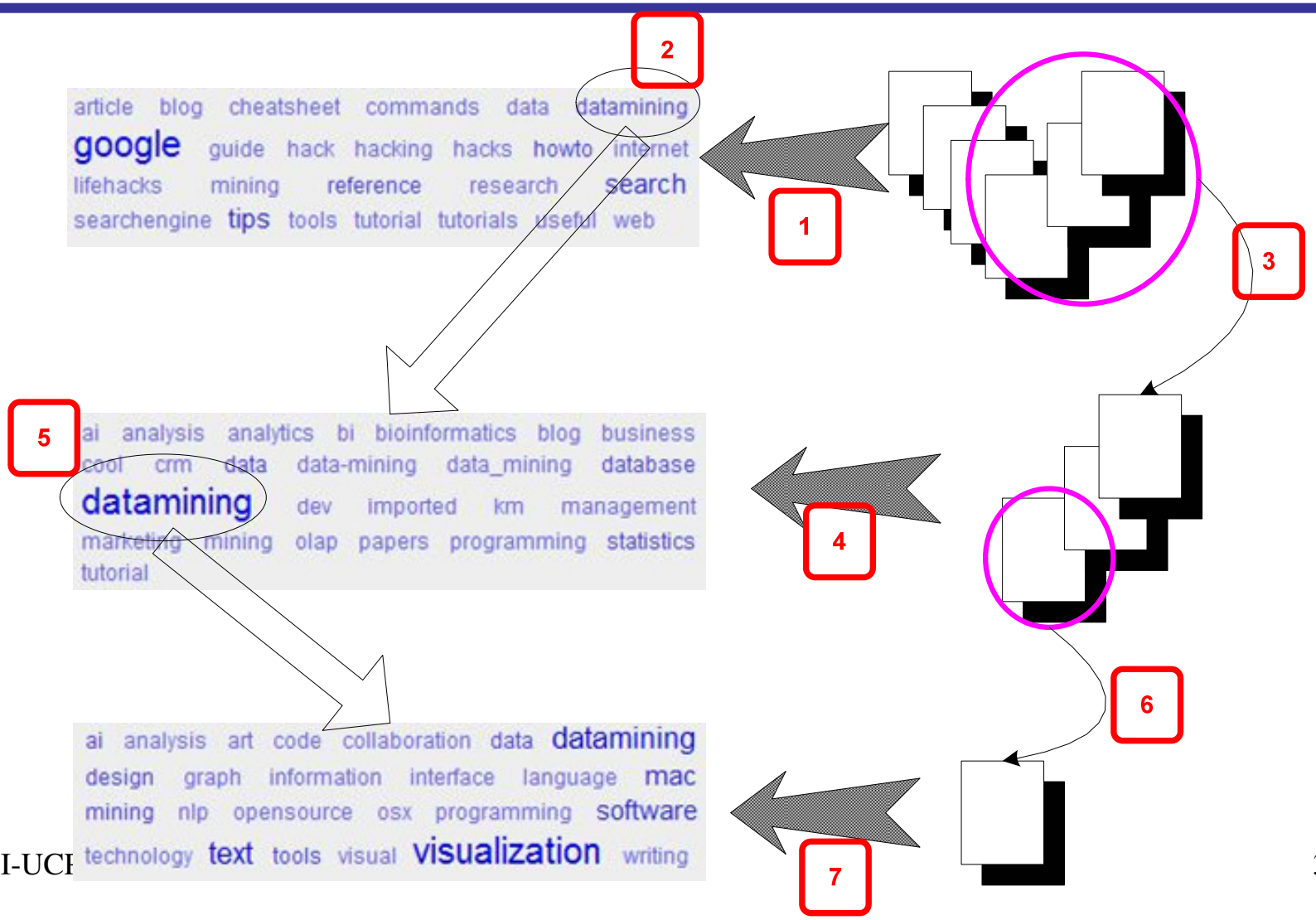
- Análise de padrões na escolha de tags pelas pessoas
  - Sites como Delicious, Flicker, ...
  - Padrões: uso de palavras do texto
    - Palavras do título: 41%
    - Uso de palavras da 1a frase: 37%
    - Uso de palavras do 1o parágrafo: 42%
    - Palavras entre as 10 mais freqüentes: 38%
  - Conclusão:
    - Pessoas costumam utilizar palavras do próprio texto
    - Palavras mais usadas tidas como melhores descritores (Inteligência Coletiva)
    - Escolha automática de tags para substituir trabalho manual (alguns sites de notícias usam editores humanos)

Tag Clouds

article blog cheatsheet commands data datamining
google guide hack hacking hacks howto internet
lifehacks mining reference research search
searchengine tips tools tutorial tutorials useful web



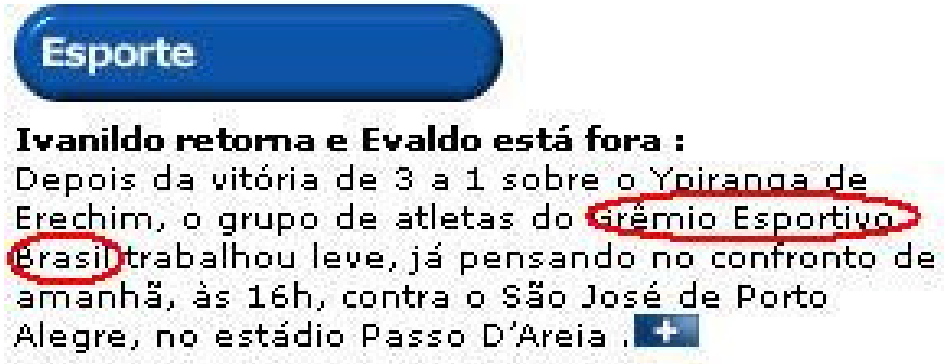
Projeto Tag CloudYAH
Hierarquia de Tag Clouds



Georreferenciamento de Notícias

- Objetivo:
  localizar geograficamente notícias (ou outros tipos de textos)
  pela análise do texto

- Aplicações:
  resultados de busca personalizados
  melhor visualização
  buscas específicas



Georreferenciamento de Notícias

- Ontologia Geográfica:
  - Entidades geográficas (lugares) + coordenadas geográficas
  - Análise de nomes próprios
  - Análise de contexto (ex: Rio de Janeiro)
  - Desambigüação (ex: Ipanema)

Pelotas=(Esporte Clube Pelotas[Lobão], Grêmio Atlético Farroupilha, Grêmio Esportivo Brasil[Xavante], Diário Popular, Canal São Gonçalo,Adolfo Antônio Fetter Jr.[Prefeito], Laranjal, Fenadoce, Universidade Católica de Pelotas[UCPEL], Universidade Federal de Pelotas[UFPEL])

Georreferenciamento de Notícias

- Construção da ontologia
  - Especialistas
  - Análise de relações entre palavras
    - Ex: Copacabana X Rio de Janeiro, Maluf X São Paulo
  - Buscas em sites como Wikipedia

FIM

http://gpsi.ucpel.tche.br

Financiamento
e apoio:

Conselho Nacional de Desenvolvimento
Científico e Tecnológico

Fundação de Amparo à Pesquisa
do Estado do Rio Grande do Sul

Universidade Católica de Pelotas