

Aprendizagem de Maquina Aplicada a Sistemas de Recomendacao

Renata S. S. Guizzardi
Depto. de Sistemas de Raciocinio
Automatizado, ITC-IRST, Trento, Italy



O que e Aprendizagem de Maquina?

Sub-area da Inteligencia Artificial que se concentra no desenvolvimento de algoritmos que permitam o computador a aprender (Wikipedia)



Areas de Aplicacao

- Recuperacao de Informacao
 - Maquinas de Busca
 - Catalogos
- Classificacao de Sequencias de DNA
- Financas
 - Analise de Bolsas de Valores
- Robotica
 - Locomocao
- Jogos
- Processamento Visual
 - Reconhecimento de caracteres em linguagem escrita
 - Reconhecimento de imagem
- Processamento da Fala
- ...

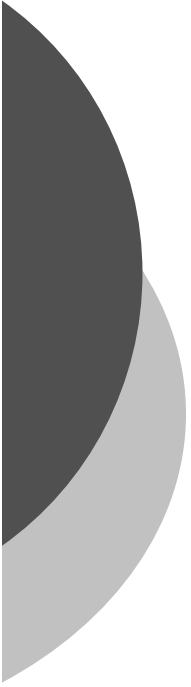
Aprendizagem de Maquina & Reconhecimento de Imagens



E. Olivetti, S. Veeramachaneni, D. Sona
ITC-IRST, Trento, Italy

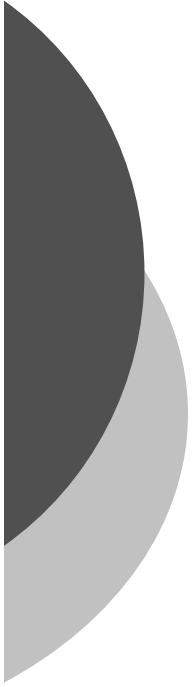
Primeiro Premio na
Competicao intitulada
*"Pittsburgh Brain
Activity
Interpretation
Competition: Inferring
Experience Based
Cognition from fMRI"* -
Human Brain Mapping
Conference (HBM),
2006.

<http://www.ebc.pitt.edu/competition.html>

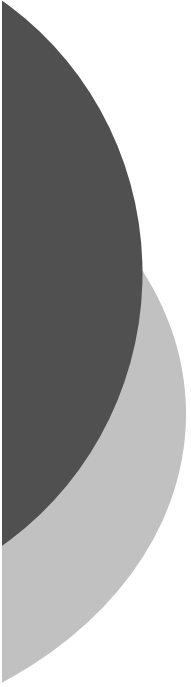


Aprendizagem de Maquina & Recuperacao de Informacoes

- *Sociedade Genealogica de Utah (GSU)*:
 - fundada em 1894,
 - para recolher dados referentes a genealogia.
- A GSU investe em treinamento de pessoal especializado para diversas tarefas, tais como:
 - processamento de imagem e de metadados;
 - armazenamento e preservacao;
 - indexacao e catalogacao;
 - acesso e distribuicao.
- Hoje, *A Biblioteca de Historia Familiar* contem:
 - mais de 2 milhoes de microfilmes;
 - 400.000 micro-fichas;
 - 300.000 livros.



O que é aprender?



Elementos de um Problema de Aprendizagem de Maquina

- Classe de tarefas (T)
- Medida de performance a ser melhorada (P)
- Fonte de experiencia (conhecimento) (E)



Exemplo (1/2)

- Sistema para adivinhar em que lugar um determinado time vai ficar na proxima Copa do Mundo.
 - T: dado um time, determinar se ele vai sair da Copa na primeira fase, nas oitavas ou nas quartas, ou se vai ficar em primeiro, segundo, terceiro e quarto lugar.
 - P: medida de acertos dada a proxima Copa
 - E: colocacao dos times nas ultimas copas

Exemplo (2/2)

T:

Time	Colocação em 2010
Brasil	?
França	?
Argentina	?
...	

E:

Time	Colocação				
	2006	2002	1998	1994	...
Brasil	Oitavas	1	2	1	
França	2	Primeira etapa	1	-	
Argentina	Oitavas	Primeira etapa	<u>Quartas</u>	Oitavas	
...					

P: quantos acertos na tabela T dada a classificacao real em 2010



Classificacao de Metodos

- **Aprendizagem Supervisionada:** o sistema gera uma saída a partir da entrada considerando determinados parametros.
Ex. Classificacao
- **Aprendizagem Nao-supervisionada:** o sistema gera uma saída a partir da entrada, porem sem contar com parametro algum.
Ex. *Clustering*
- **Aprendizagem por Reforco:** o sistema aprende a partir da observacao das consequencias de suas acoes no ambiente (geralmente a partir de feedback).
Ex. Planejamento



Principais Tecnicas de Recomendacao

- Baseadas em Conteudo: verifica similaridades entre os itens de informacao.
- Baseadas em Colaboracao: verifica similaridades entre usuarios (perfil).
- Hbridas:
 - combina as duas anteriores e outras tecnicas.
 - novidades: combinar características cognitivas (ex. confiança, função cognitiva)



Topicos de Pesquisa em Sistemas de Recomendacoes

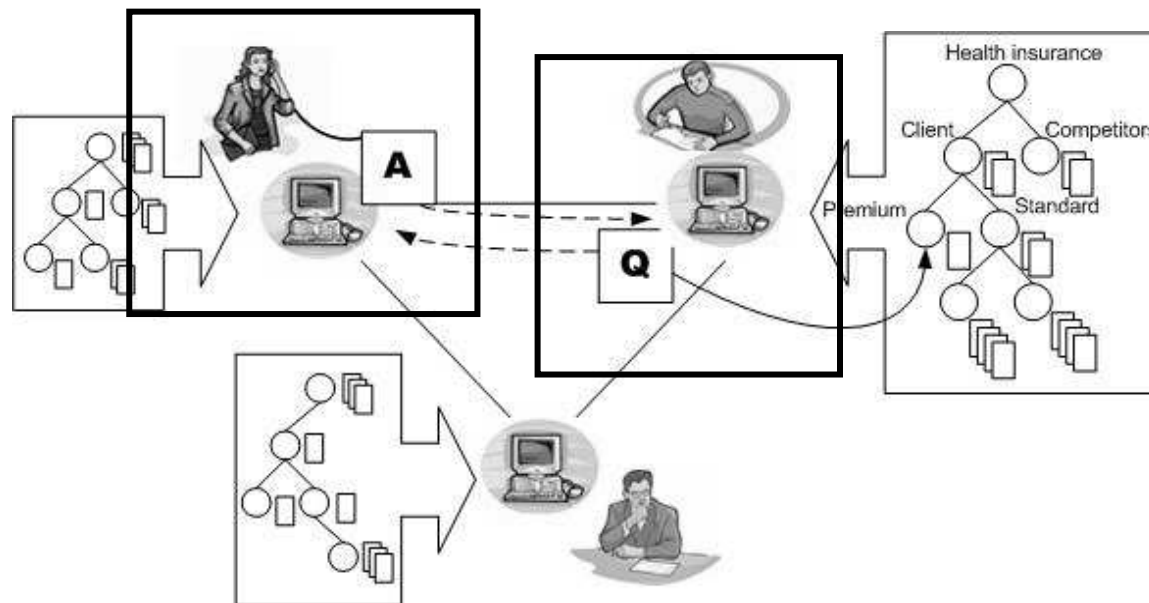
- Modelagem da Informacao
- Recomendacao a partir de classificacao e clustering
- Perfis/Modelos de usuarios
- Bootstrap
- ...



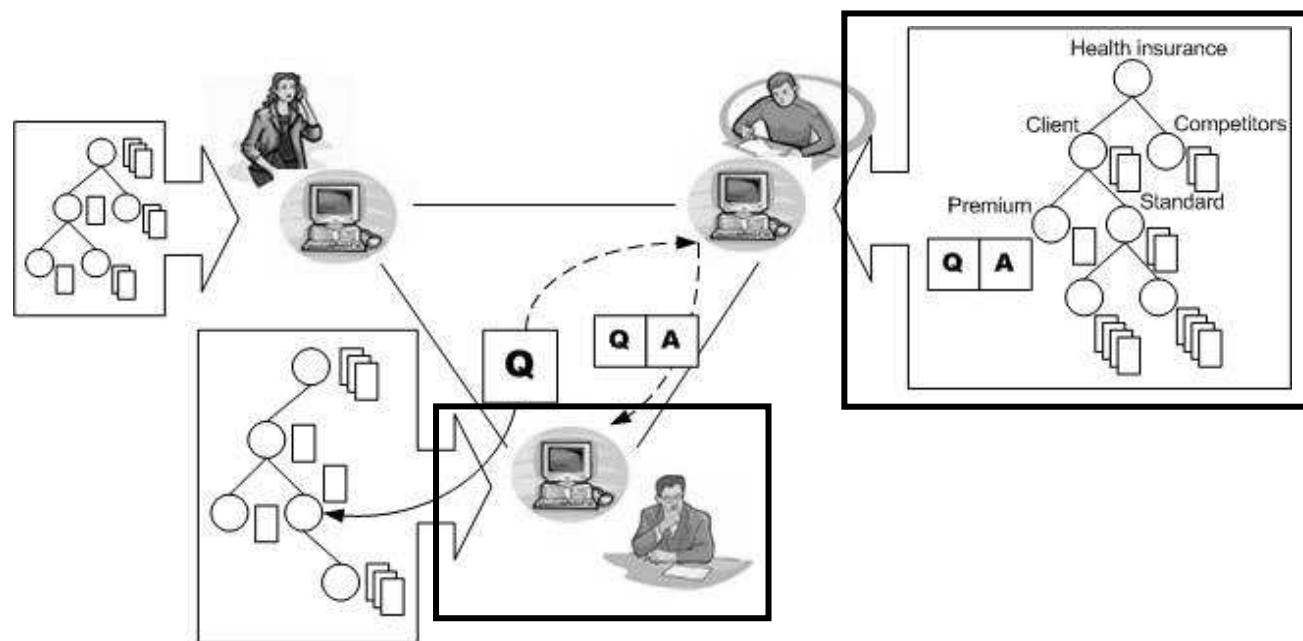
KARe: Knowledgeable Agent for Recommendations

- Objetivo:
 - suporte ao compartilhamento de conhecimento entre membros de uma organizacao
- Sistema de recomendacao de arquivos textuais – hibrido, baseado em:
 - conteudo
 - aspectos cognitivos do usuario
- Caracteristicas principais:
 - usuario *autonomo* na organizacao e troca de conhecimento (peer-to-peer);
 - recomendacao baseada em perguntas e respostas;
 - algoritmo de recomendacao baseado no uso de informacoes taxonomicas e modelagem vetorial de dados.

Funcionamento do KAREe (1/2)



Funcionamento do KARE (2/2)

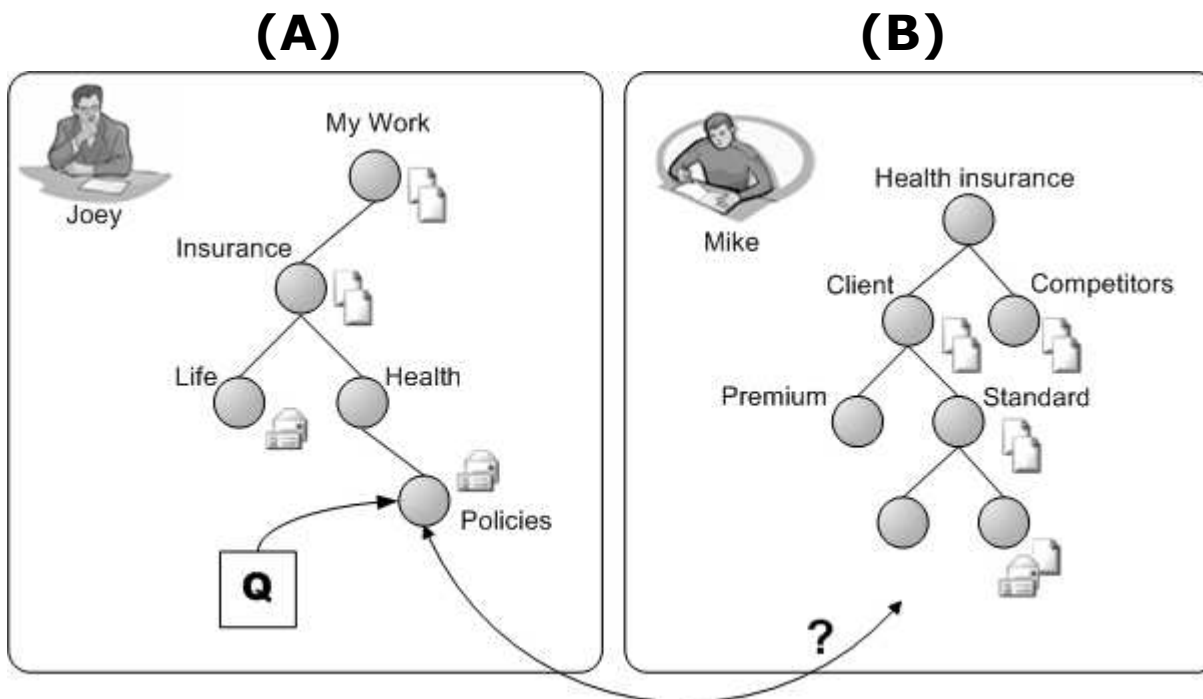




Elementos de Aprendizagem de Maquina em KARE

- T: recuperar uma resposta previamente armazenada no sistema a uma pergunta de entrada.
- P: medida de satisfacao fornecida pelo usuario (feedback)
- E: taxonomia, documentos classificados, perguntas-respostas anteriores

T – Primeiro passo

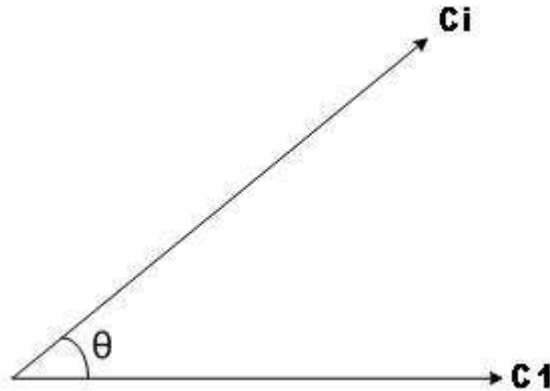


Dada a questão Q classificada no conceito C1 da taxonomia A, qual o conceito C2 da taxonomia B em que a resposta de Q se encontra?

Ou seja, $C2 \approx C1$

Busca pelo Conceito Equivalente

- Conceitos são modelados como vetores multidimensionais.
- Similaridade calculada a partir do cosseno entre os vetores C_1 e C_i .



Dimensoes dos Vetores

- Vocabulario: Todas as palavras contidas nos documentos de uma taxonomia sao computadas em um vetor.

Game	soccer	Brazil	car	Ferrari	go
------	--------	--------	-----	---------	----

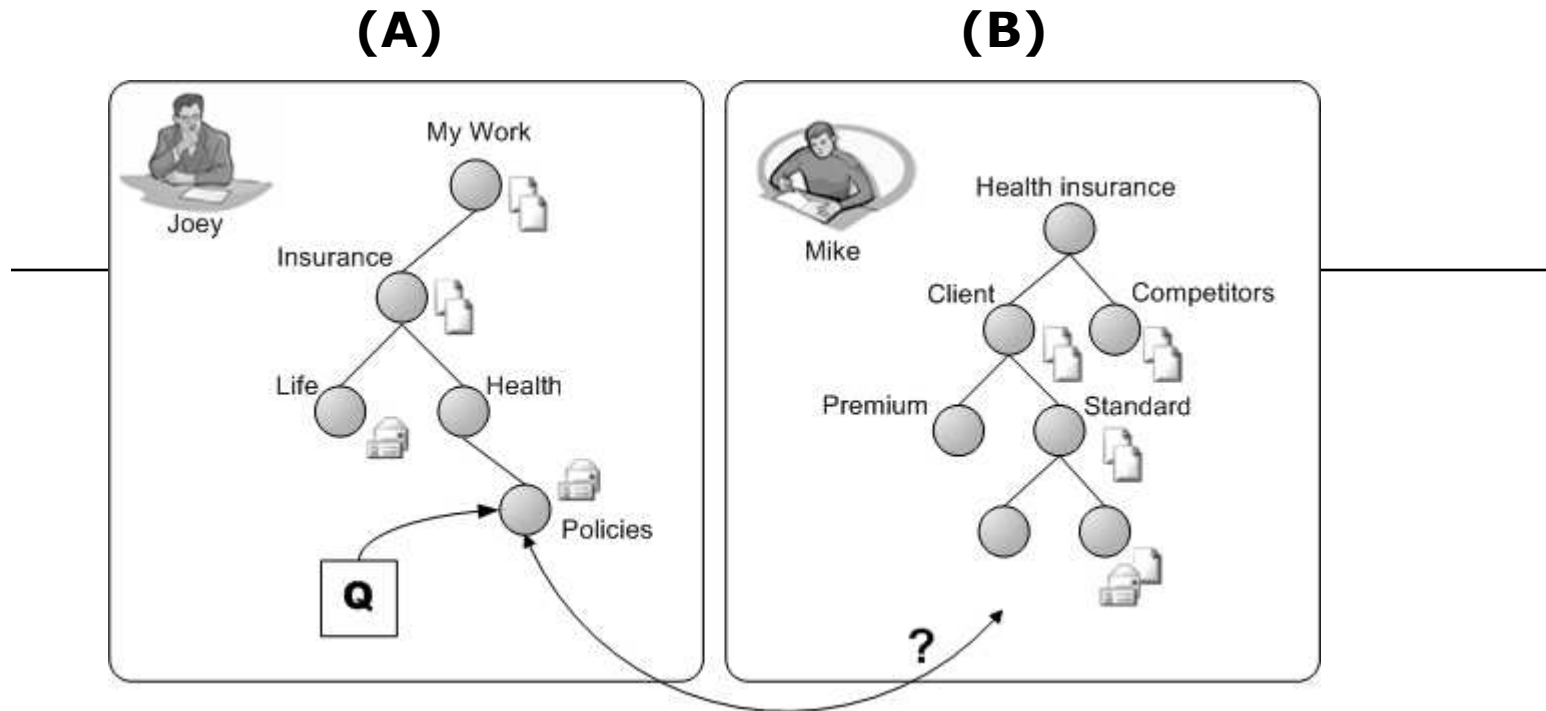
- Vetor do Conceito: cada conceito e representado por um vetor em que cada posicao indica o *peso* da palavra naquele conceito.

VC1: conceito contem palavras "game" e "go"

1	0	0	0	0	1
---	---	---	---	---	---

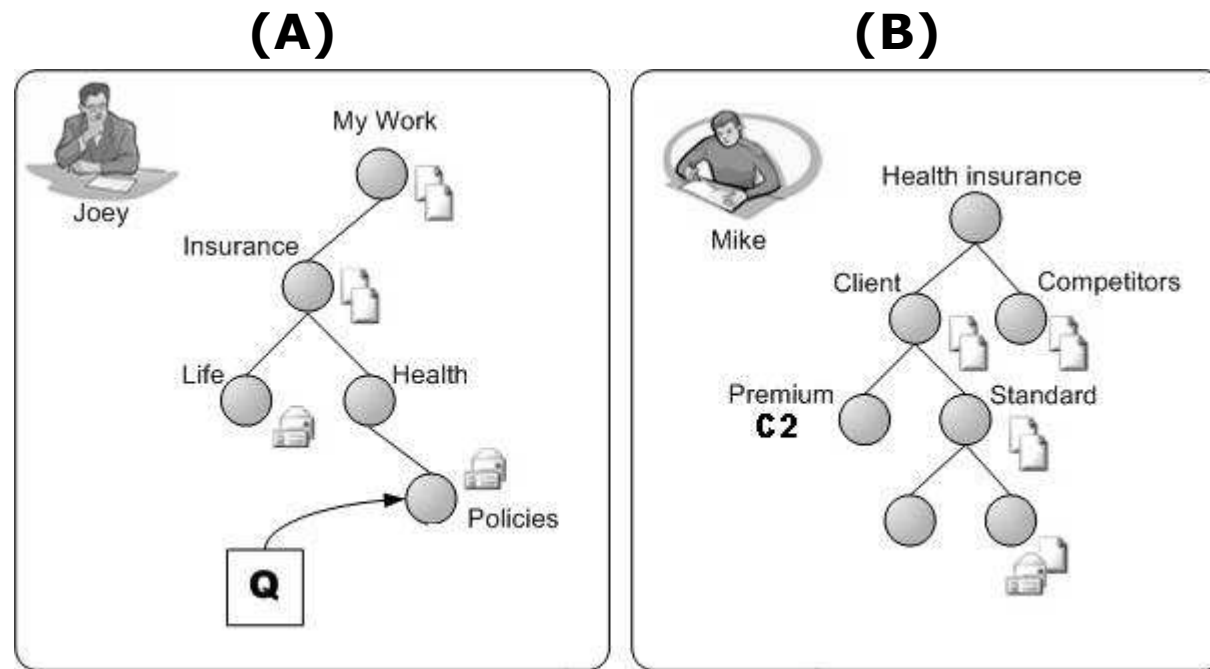
VCi: conceito contem palavras "game", "soccer e "car"

1	1	0	1	0	0
---	---	---	---	---	---



- Ci e representado por dois vetores:
 - Um contendo as palavras correspondentes ao nome dos conceitos e seus antepassados.
 - Um contendo as palavras contidas nos documentos classificados pelo conceito.
- Como vocabulário $A \neq$ vocabulário B , então há uma *tradução*, representando os vetores de C_1 de acordo com o vocabulário B .

T – Segundo passo



Dado C2, encontrar a resposta a pergunta Q dentre os itens de informação (documentos e perguntas/respostas) classificados em C2.



Busca pela resposta de Q

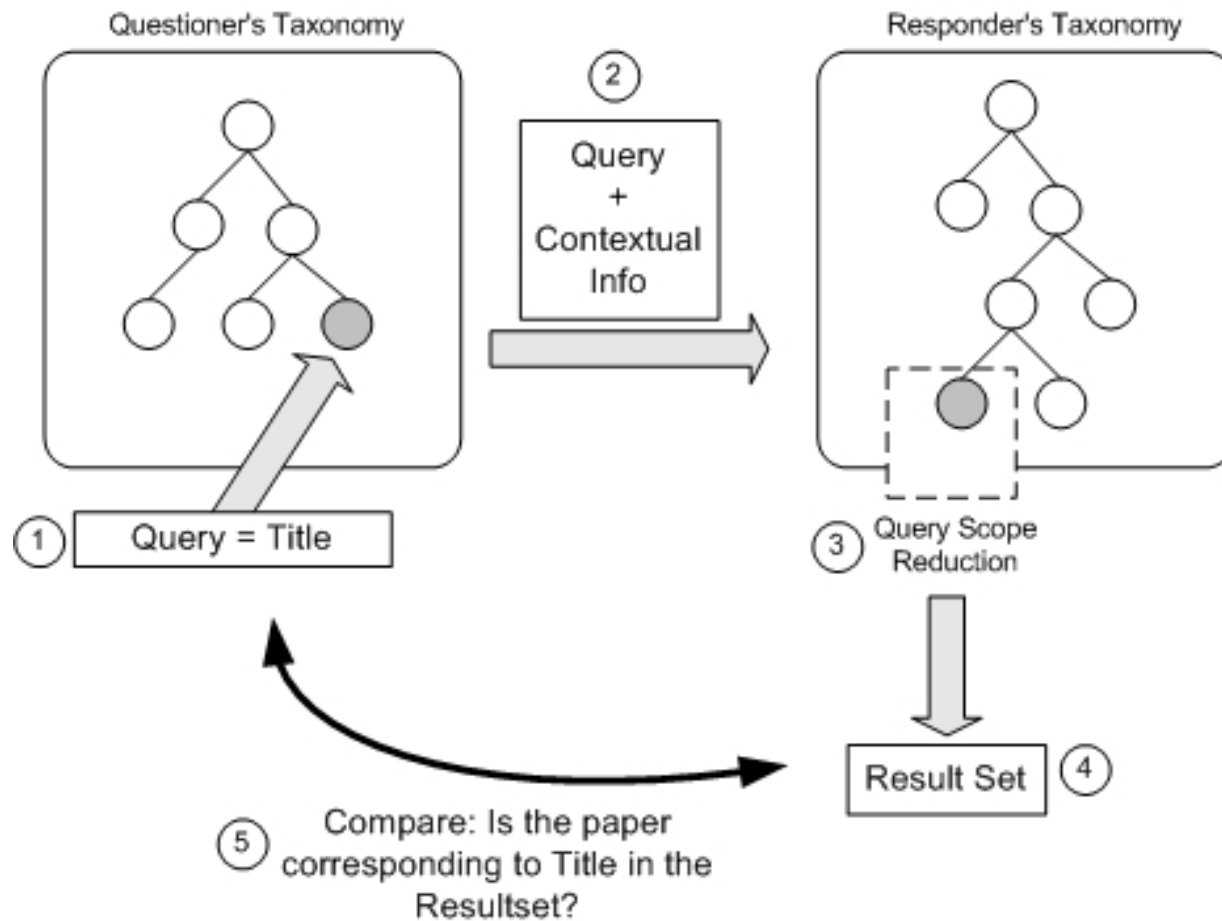
- Q:
 - vetor de palavras contidas no texto da pergunta.
- Itens de Informacao classificados no conceito:
 - vetor de palavras contidas no proprio item de informacao.
- Similaridade
 - ainda calculada usando o cosseno entre os vetores (qto menor o cosseno, mais proximos os vetores).



Outros Detalhes

- O texto de perguntas e de itens de informacao sao pre-processados:
 - reducao de *stopword*: considera-se apenas *palavras-chave* na representacao do vetor;
 - *stemming*: reduz-se os termos a sua raiz (ex. Caminhar, caminhando, caminhei sao consideradas a mesma palavra)
- Permite-se encontrar mais de um conceito equivalente, dando mais flexibilidade a busca.
- A resposta a pergunta e feita a partir do ranqueamento de itens de informacao a partir da similaridade.

Experimento (1/2)





Experimento (2/2)

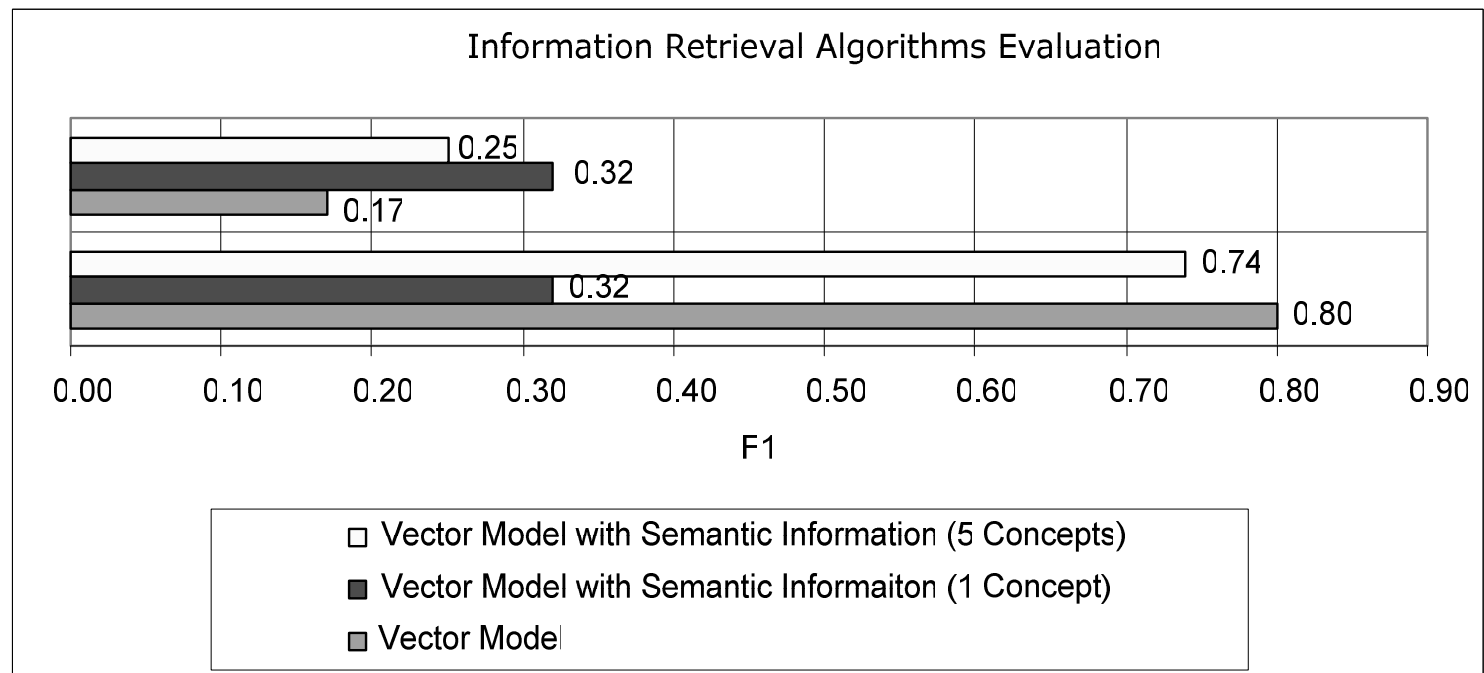
- A pergunta Q é simulada como o título de um artigo científico
- A resposta é o corpo do artigo (sem o título para evitar *bias*)
- O experimento mede:
 - Se o conceito C2 é corretamente encontrado (C2 é o conceito que classifica o artigo correspondente ao título);
 - Se o item que corresponde à resposta à pergunta Q (no caso, o artigo pertencente ao título) é corretamente encontrado.



Resultados (1/2)

- Nosso algoritmo encontra, em media, 5 vezes mais o conceito correto que a abordagem classica;
- Geralmente, a abordagem classica encontra o documento especifico mais vezes;
- Se varios conceitos equivalentes sao considerados ao inves de 1 apenas, a performance do nosso algoritmo para encontrar o documento especifico aumenta.

Resultados (2/2)



Prototipos

Desktop

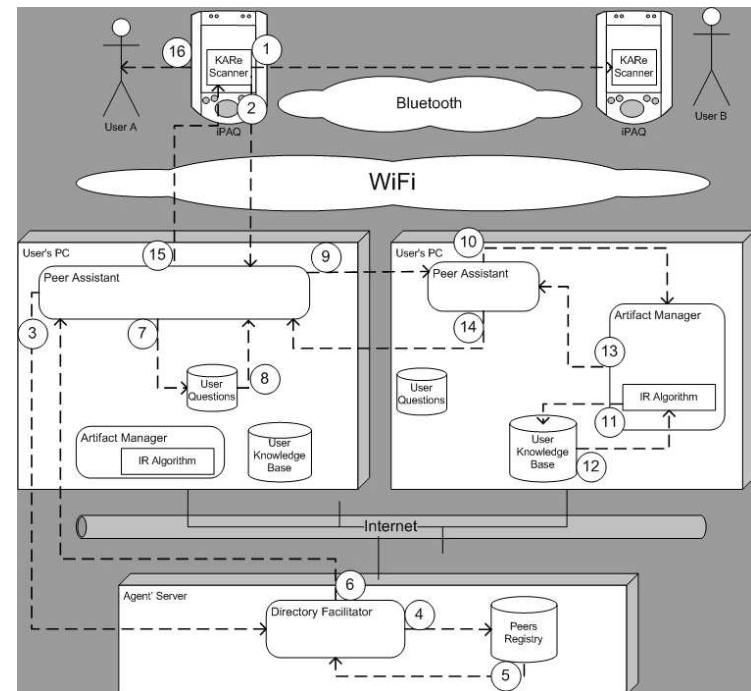
KARe - Knowledgeable Agent for Recommendations

Understanding Process and the Quest for Deeper Questions in Software Engineering Research

Search

Peer	Concept	Artifact	Similarity
Pablo	Software Engineering	Understanding Process and the Quest for Deeper Questions in Software Engineering Research	0.432602673...
Pablo	Software Engineering	SOFTWARE ENGINEERING - AN ECONOMIC PERSPECTIVE	0.109004501...
Pablo	Software Engineering	Information Systems Security Engineering: A Critical Component of the Systems Engineering L...	0.100261555...
Pablo	Software Engineering	THE POLITICS OF SOFTWARE ENGINEERING	0.090990334...
Pablo	Software Engineering	EDUCATIONAL EXPERIENCES IN INDUSTRIAL SOFTWARE ENGINEERING	0.085918970...
Pablo	Software Engineering	THE IEEE SOFTWARE ENGINEERING STANDARDS PROCESS	0.085243016...
Pablo	Software Engineering	How to Manage Your Software Product Life Cycle with MAUI	0.082061395...
Pablo	Software Engineering	PROGRESSING FROM STUDENT TO PROFESSIONAL: THE IMPORTANCE AND CHALLENGE	0.05349646...
Pablo	Software Engineering	A PROJECT ORIENTED COURSE ON SOFTWARE ENGINEERING	0.051162801...
Pablo	Software Engineering	Experimenting with Pair Programming in the Classroom	0.050825767...

Handheld





Conclusões

- A aprendizagem de máquina é uma das áreas mais pesquisadas da IA, no passado, presente e futuro.
- Varias áreas de aplicação importantes, dentre elas:
 - máquinas de busca e
 - sistemas de recomendação.
- Quanto a recomendação, a aprendizagem de máquina pode ser aplicada:
 - em problemas de classificação ou clustering de informação
 - na criação de perfis de usuário
 - Para solucionar limitações de *bootstrapping*

Google™

Akwan é agora parte do Google Brasil.
Saiba mais.

- **Pesquisa de web específica ao país:** Se você está interessado em buscadores que são específicos a um país, tente Google Ferramentas de idiomas para visitar o site do Google no seu domínio local, como o Google Brasil.
- **Pesquisa de web no seu site:** Se você está interessado em providenciar pesquisa de web no seu site, aprenda sobre Google AdSense para pesquisas.

"A partir de agora, a Akwan, que possui sede em Belo Horizonte, vai se tornar o Centro de Pesquisa e Desenvolvimento (P&D) do Google na América Latina. O objetivo da empresa norte-americana é aumentar sua atuação no continente, através do desenvolvimento de tecnologia de ponta e recrutamento de mão-de-obra qualificada em toda a região."

(texto retirado de site de notícias na Web)