# IMPLEMENTAÇÃO DE UM *CRAWLER* INCREMENTAL DISTRIBUÍDO: UM SISTEMA DE BUSCA NA *WEB*

Tiago Roberto Fischer

Alexander Roberto Valdameri - Orientador

UNIVERSIDADE REGIONAL DE BLUMENAU

# Roteiro

- Introdução
- Objetivos
- Fundamentação
- Desenvolvimento
- Conclusão e extensões

# Introdução

- Internet (*web*)
- Sistemas de Busca
- *Crawlers*

# Objetivos

- Possibilitar a busca de páginas na *web*
- Possibilitar o armazenamento das páginas localmente de forma indexada
- Prover comunicação entre todos os *crawlers*
- Manter as páginas coletadas atualizadas
- Disponibilizar uma interface para busca básica de conteúdos nos *crawlers* de forma transparente

# MYSQL

- Acesso seqüencial (*MyISAM*)
- Índices únicos
- *Full-text search* (índice invertido)
- Armazenamento e recuperação

# JAVA

- Multi-plataforma
- Orientado à objetos
- Classes (*syncronized*)
- Processos (*thread*)

# PHP

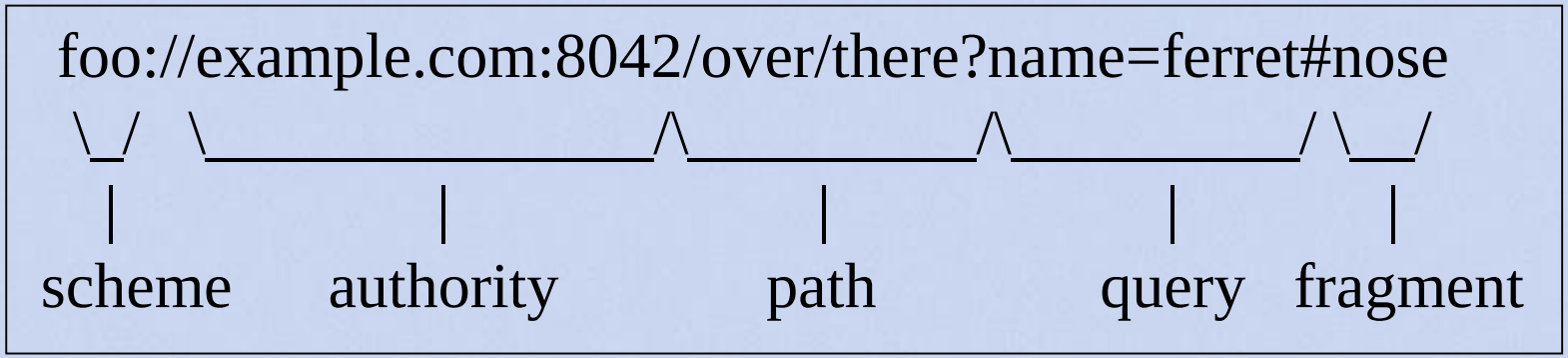
- Programação para *web*
- Interação com o MYSQL
- Facilidade

# Sistemas de Busca

- *Crawlers*
- *URL server*
- *Sorter/indexer*

- *Ranking*

# URI/URL



# Métodos de busca e métricas

- FIFO
- LIFO
- *Backlink count*

# *Crawler*

- Incremental
- Periódico

# Sistemas distribuídos

- Escalabilidade
- Transparência

# Requisitos do problema (*crawler*)

- Buscar páginas na *web* e armazená-las
- Processar páginas à procura de novas URL
- Prover comunicação com todos os *crawlers* em funcionamento
- Manter todo o banco de dados (coleção global de páginas) de forma eficaz, não tendo páginas duplicadas em diferentes *crawlers*
- Manter atualizada a sua coleção local de páginas
- Processar informações pedidas pelo *web search* na sua coleção local de páginas

# Requisitos do problema (*web search*)

- Disponibilizar uma interface para a busca de informações
- Comunicar-se com os *crawlers* para efetuar a busca de forma transparente para o usuário

# Requisitos do problema (*sorter/indexer*)

- Manter organizada fisicamente todos as páginas
- Manter indexadas todas as páginas

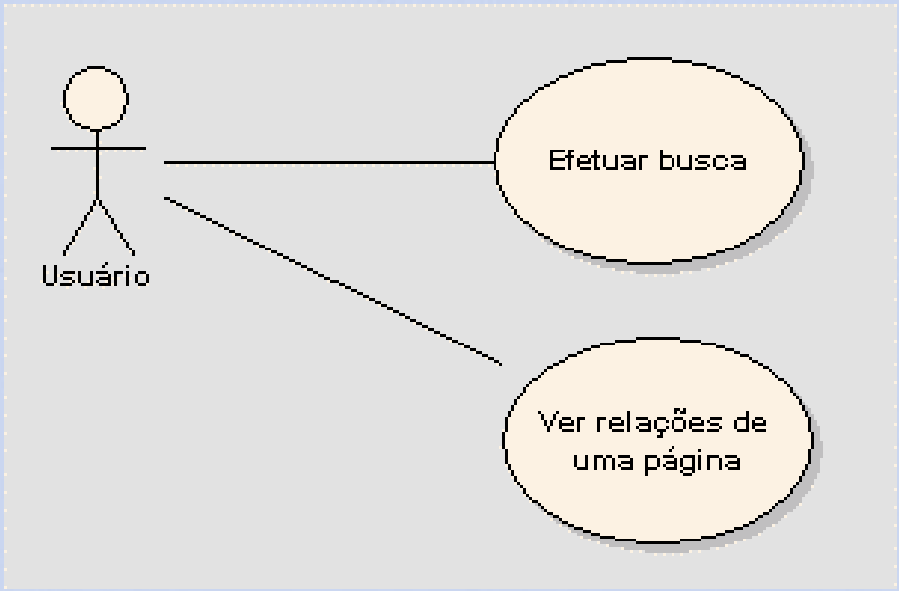
# Requisitos do problema (servidor de URL)

- Manter uma relação de todas URL
- Manter o tipo (MIME-TYPE) de cada URL e o *crawler* a qual a URL pertence
- Fazer a distribuição de *links*

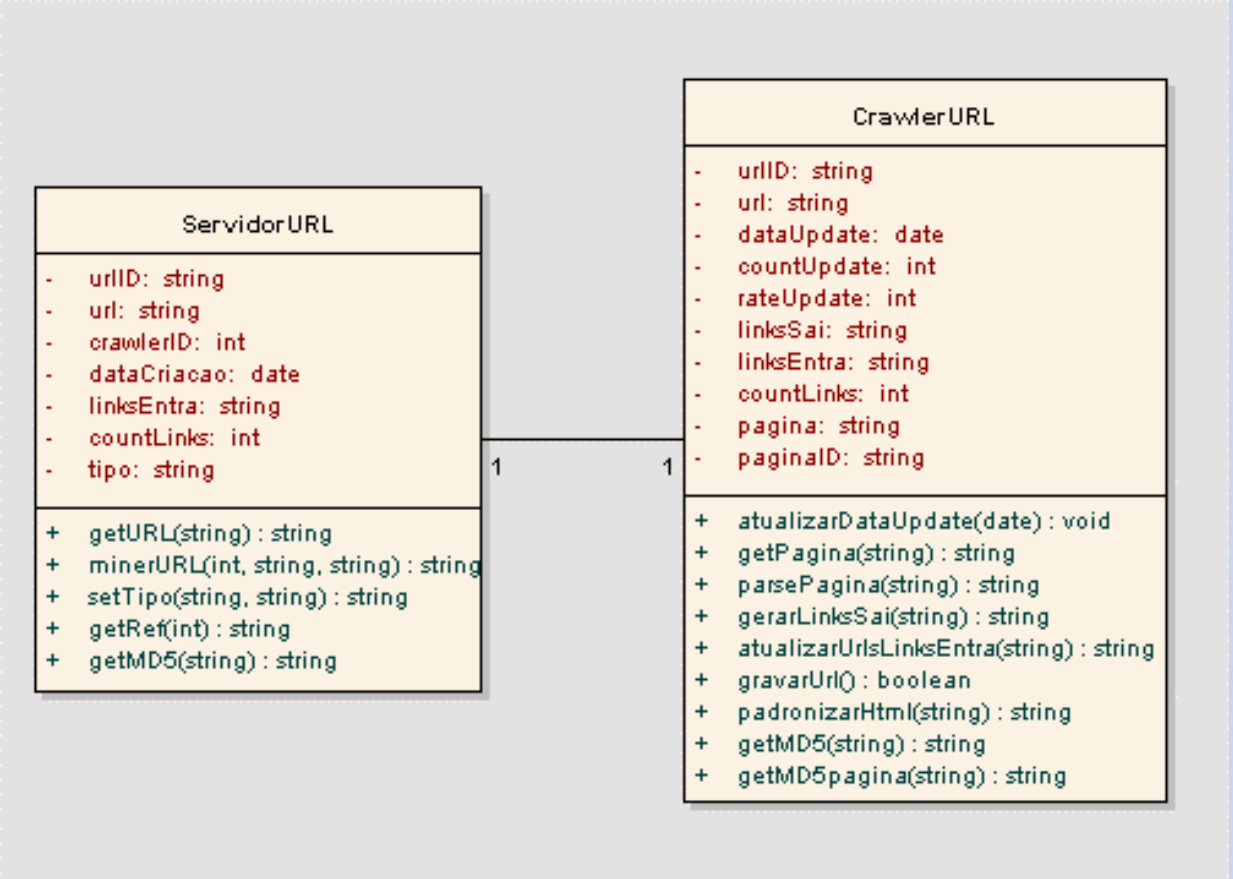
# Especificação

- Casos de uso (Enterprise Architect)
- Diagrama de classes (Enterprise Architect)
- MER (DBDesign)

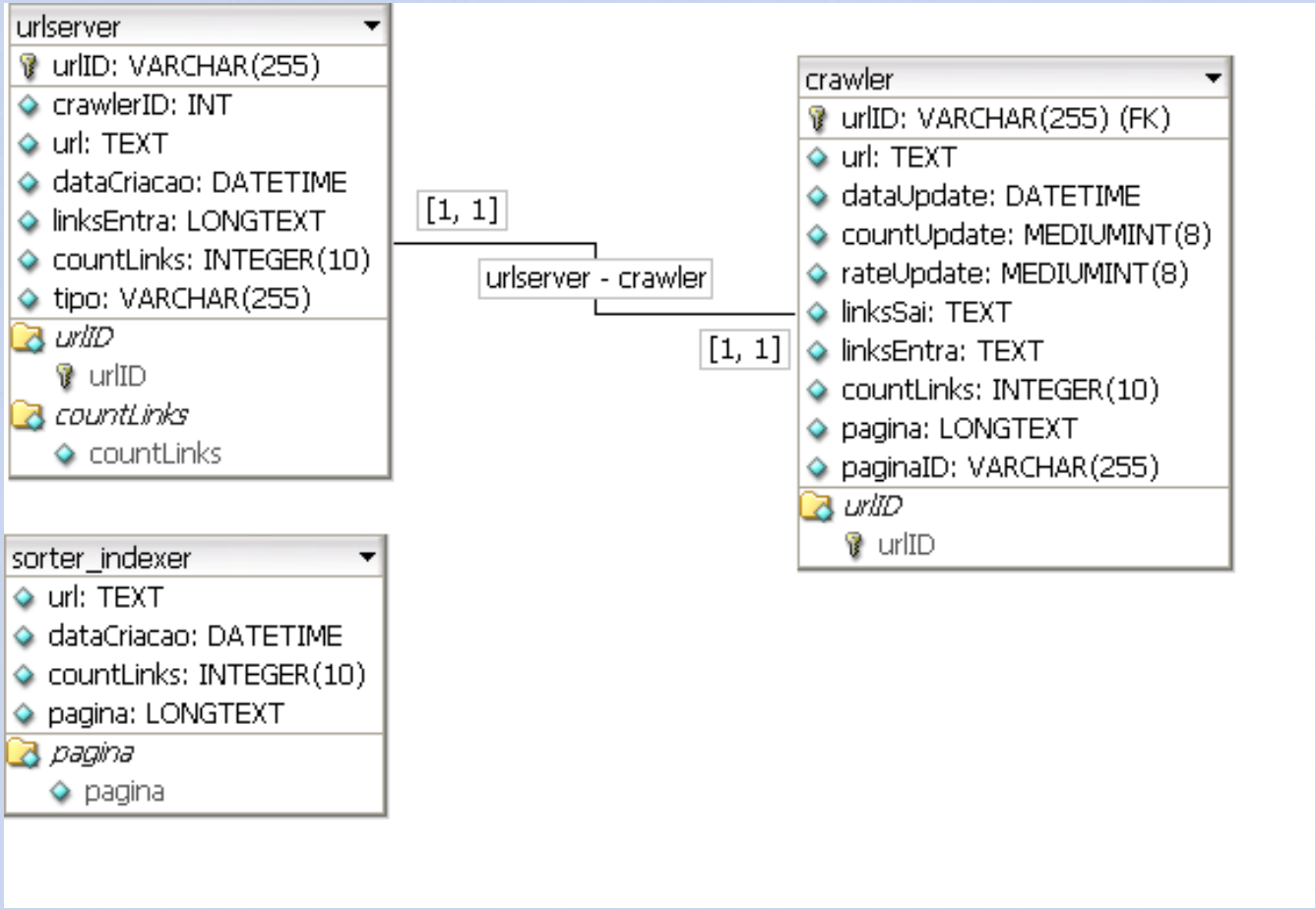
# Casos de uso



# Diagrama de classes



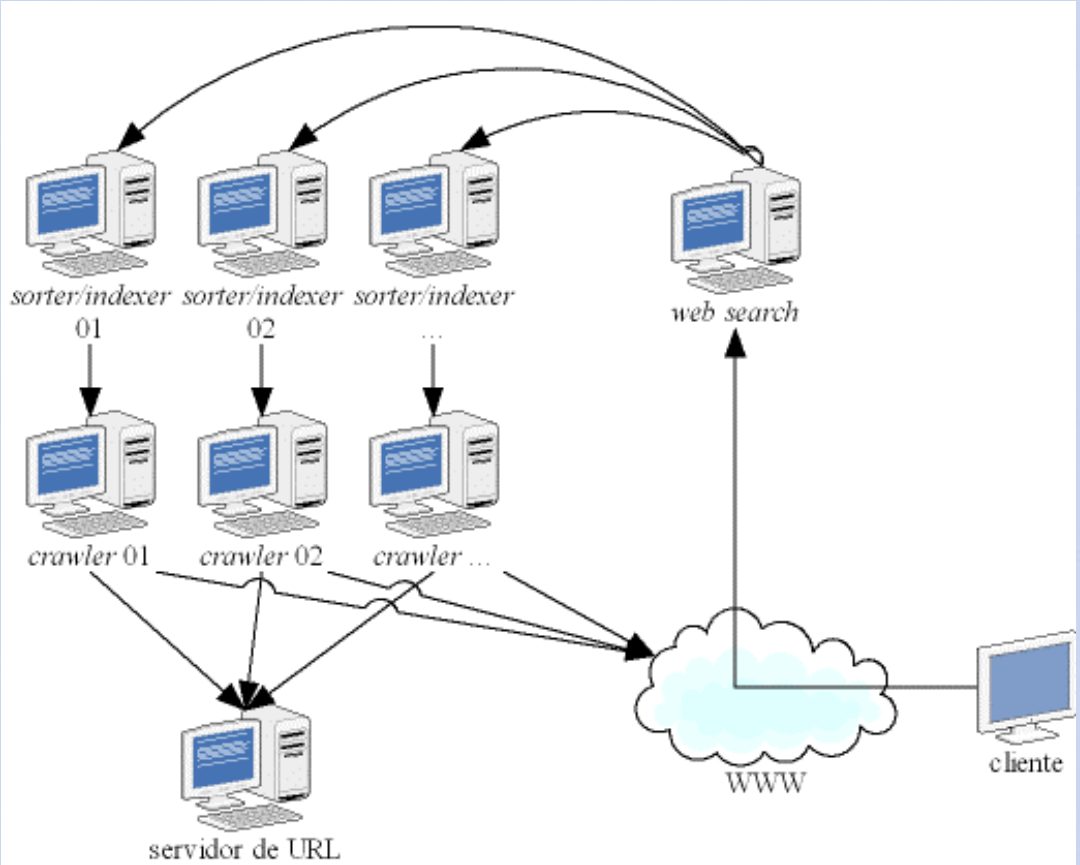
# Modelo entidade-relacionamento



# Ferramentas Utilizadas

- MySQL
- Netbeans 5.0 (Java)
- PHP

# Visão Geral



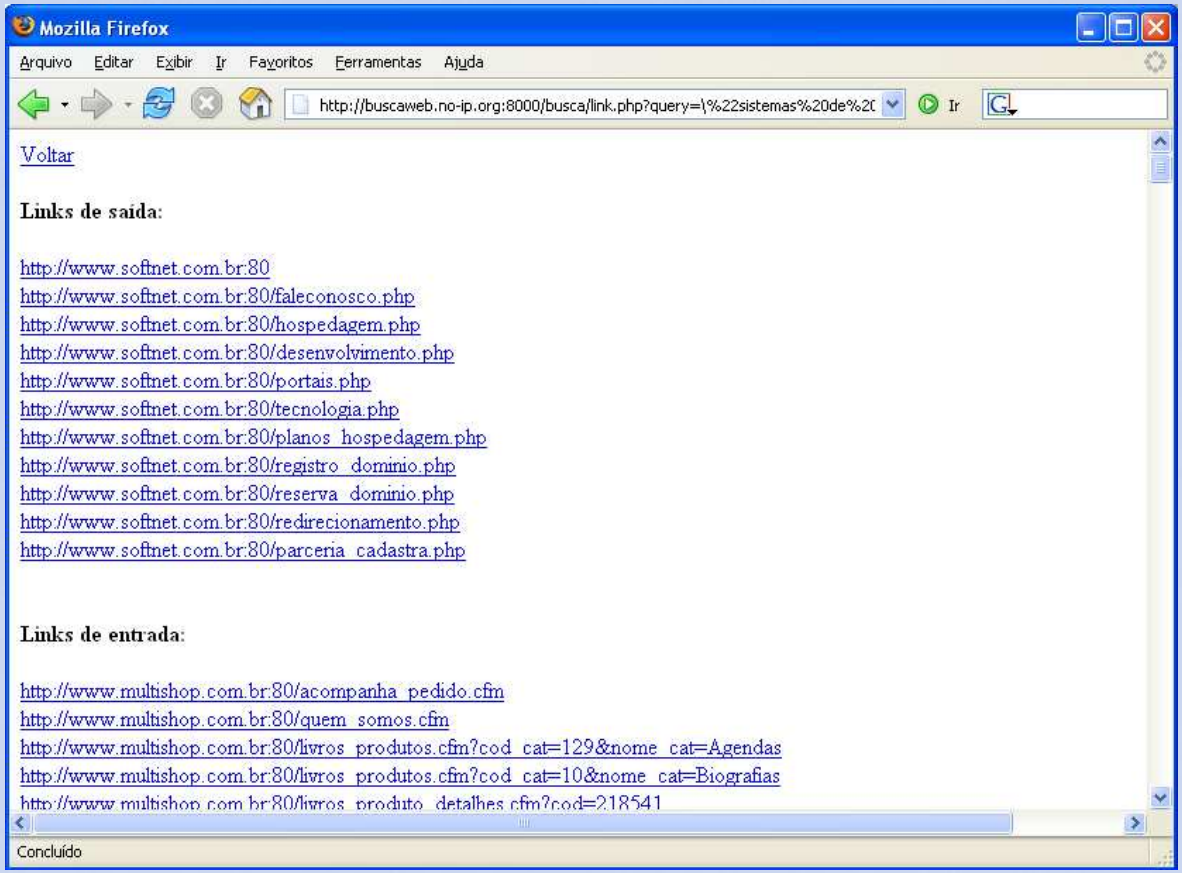
# Operacionalidade da Implementação

# Operacionalidade da Implementação

| busca | tempo de retorno (segundos) | tempo de retorno da segunda busca (segundos) |
|---|---|---|
| busca por qualquer palavra: **café carro bolo** | 0.1969 | 0.0020 |
| busca por qualquer palavra: **cachorro gato** | 0.1900 | 0.0019 |
| busca por todas as palavras: **+informatica +banco +dados** | 0.2037 | 0.0021 |
| busca por todas as palavras: **+aparelho +auditivo** | 0.3235 | 0.0019 |
| busca por uma palavra sim e outra não: **+navegador –firefox** | 0.1052 | 0.0019 |
| busca por uma palavra sim e outra não: **+celular -compra** | 0.2185 | 0.0019 |
| busca por frase: **“sistema de busca”** | 0.1076 | 0.0021 |
| busca por frase: **“quero café”** | 17.86 | 0.0012 |

# Operacionalidade da Implementação

# Operacionalidade da Implementação



# Resultados e discussão

- Ponto de gargalo (servidor de URL)
- Identificação de possíveis otimizações
- Fator de profundidade da busca

# Conclusão

- Objetivo de buscar e manter o relacionamento das páginas foi alcançado
- Mysql supriu as necessidades de armazenamento e indexação
- Exemplo das necessidades e da criação (inicio à fim) de um sistema de busca na *web*

# Extensões

- distribuir o servidor de URL, tornando-o escalável
- aprimorar o ranking, possibilitando um retorno mais próximo do desejo do usuário
- possibilitar que o *sorter/indexer* seja executado de maneira paralela
- separar partes do HTML para terem mais importância em uma busca (exemplo: título da página)
- otimização (informações em memória, *cache*, etc)

# Relevância pessoal

- Consolidação do conhecimento aprendido durante o curso
- Ampliação dos conhecimentos
- Desafios superados