

Relatório Técnico do projecto ARIADNE

Praxis XXI

Interface de utilizador do NewsSearch

Carlos Correia
Norman Noronha
Daniel Gomes

Junho de 2000

Índice

1. INTRODUÇÃO	3
1.1 MOTIVAÇÃO.....	3
1.2 PROPOSTO	3
1.3 MÉTODO.....	3
1.4 ORGANIZAÇÃO DO DOCUMENTO:.....	4
2. CONTEXTO	5
2.1 NEWSSEARCH	5
2.2 GLIMPSE.....	5
3. TRABALHO DESENVOLVIDO	7
4. ARQUITECTURA / IMPLEMENTAÇÃO.....	8
4.1 INTERFACE DE PESQUISA:	9
4.2 GERADOR DE COLECÇÕES DE TÓPICOS.....	11
5. RESULTADOS	13
6. CONCLUSÕES E TRABALHO FUTURO.....	15
6.1 TRABALHO FUTURO.....	15
7. REFERÊNCIAS.....	16

1. Introdução

1.1 Motivação

Nos últimos anos tem-se dado um grande aumento de informação noticiosa disponível na Web, no entanto, o potencial desta disponibilidade de informação normalmente não é devidamente explorado, uma vez que os motores de busca actualmente disponíveis não colectam as edições noticiosas com a frequência desejada, além disso não disponibilizam as funcionalidades de filtragem necessárias a uma busca rápida e eficiente que vá de encontro às expectativas dos utilizadores.

Perante este cenário foi desenvolvido o NewsSearch que consiste num sistema de recolha, armazenamento e classificação de informação noticiosa que permite ao utilizador pesquisar uma base de dados noticiosa actualizada, de forma simples e eficaz.

O NewsSearch além de desempenhar um papel informativo interessante na nossa vida quotidiana é também uma preciosa ferramenta de trabalho para certos sectores de actividade como o jornalismo, ensino, investigação, etc...

1.2 Proposto

Este projecto tem em vista estabelecer a concretização do projecto da cadeira de Publicação Digital, que consiste na criação de um servidor aplicacional que realize o acesso a uma base de dados de referências a artigos recolhidos da Internet.

A descrição do sistema compreende:

- Interface de Pesquisa à Base de Dados
- Gerador Automático de Relatórios
- Sistema de Pesquisa de Artigos Similares

1.3 Método

Para o desenvolvimento deste projecto baseámo-nos no Modelo de prototipagem (ver Figura 1), ajustando-o às nossas necessidades [Pressman 1997].

O recurso a este método de desenvolvimento justifica-se devido à incerteza de requisitos e à necessidade de modularização de componentes. O Gerador de Automático Relatórios é um incremento à Interface de Pesquisa e este é apenas mais um componente que depende e reflecte o desempenho global do sistema.

Uma vez que, para obtermos resultados satisfatórios no nosso projecto necessitámos de rever por várias vezes os requisitos incluindo os de fases anteriores, as revisões implícitas no modelo de prototipagem pareceram-nos adequadas às características do nosso projecto.

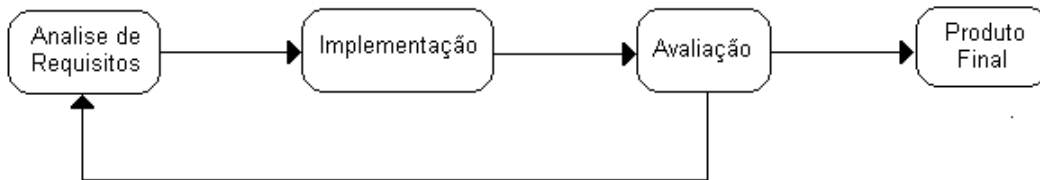


Figura 1. Metodologia de desenvolvimento do projecto.

1.4 Organização do documento:

Este documento divide-se em cinco partes que pretendem descrever as opções tomadas no decurso do desenvolvimento deste projecto.

Na primeira descrevemos o contexto inicial que serviu de base para o nosso trabalho, na segunda a arquitectura e implementação do sistema, seguida de uma descrição dos resultados obtidos e conseqüentes conclusões.

Finalizamos o documento com o trabalho futuro a ser realizado.

2. Contexto

2.1 NewsSearch

Este sistema define um conjunto de componentes de software para recuperação e classificação de informação noticiosa colecionada a partir de sites noticiosos na web (no nosso caso de sites noticiosos nacionais).

Na Figura 2 apresentamos um diagrama que descreve a arquitectura e os principais componentes do NewsSearch: Serviço de Recuperação com agentes de indexação, Serviço de Classificação, Arquivo Digital de Artigos e Portal do Sistema.

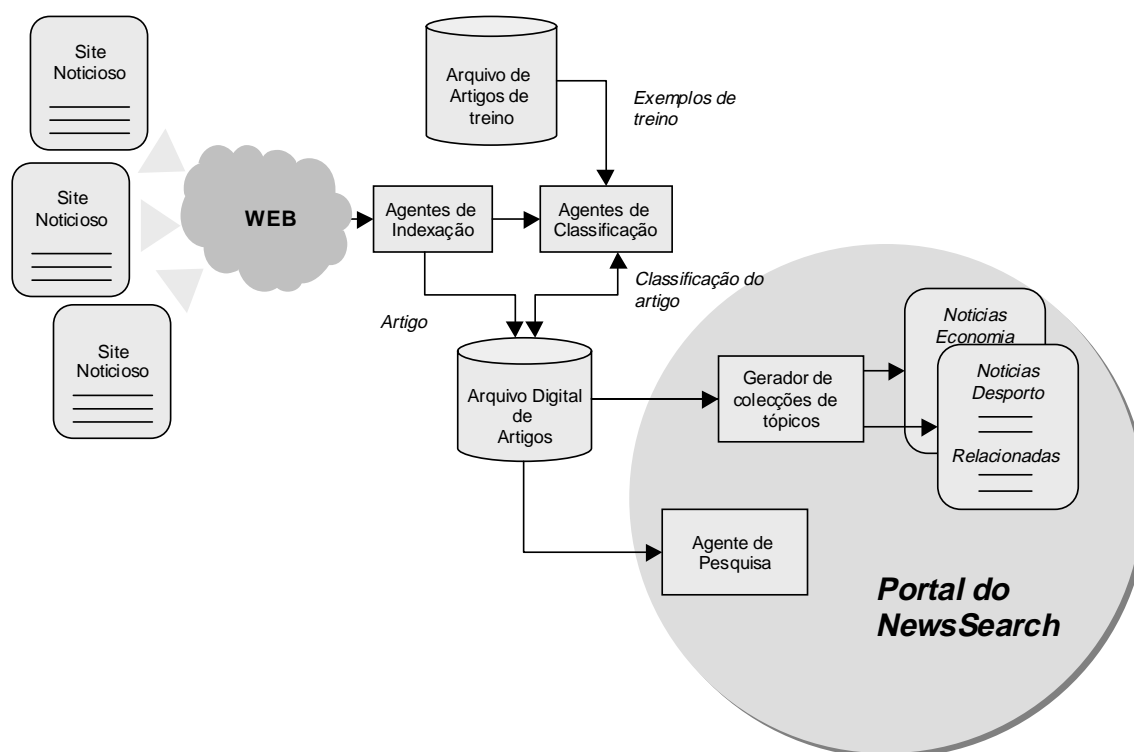


Figura 2. Arquitectura global do NewsSearch.

Uma descrição mais aprofundada acerca do NewsSearch pode ser encontrada em [Maria e Silva 2000].

O nosso trabalho incide do desenvolvimento do Portal do NewsSearch que possibilita o acesso à informação, ou seja, na Interface de Pesquisa e Geração de Colecções de Tópicos.

2.2 Glimpse

O Glimpse [Glimpse Web site] é um componente que permite a indexação e consulta eficiente de termos numa colecção de documentos. O Glimpse faz uso

de técnicas já amplamente validadas pela “Information Retrieval” [Korfage 1997].

O Glimpse é usado no nosso projecto para interagir com a biblioteca digital de informação noticiosa, suportando as operações de pesquisa de texto livre. Estas operações são suportadas recorrendo a uma estrutura de indexação, e através do uso de expressões regulares, expressões booleanas e pesquisas por aproximação.

Pode-se subdividir este sistema em dois componentes distintos:

- **Indexador** (GlimpseIndex) – cria tabelas de localização de palavras nos documentos
- **Servidor de Pesquisas** (GlimpseServer) – recorrendo às estruturas de localização responde às pesquisas que são colocadas através da Interface de Pesquisa.

3. Trabalho desenvolvido

Na análise preliminar do projecto apresentámos o seguinte planeamento:

Fase	Descrição	Prazo
0	Fase Inicial: - Configuração do ambiente - Análise de documentação - Integração no projecto	15/04/2000
1	Interface de Pesquisa	05/05/2000
2	Gerador Automático de Relatórios	15/05/2000
3	Sistema de pesquisa de artigos similares	25/05/2000
4	Documentação e testes	29/05/2000

Infelizmente não foi possível realizar todas as fases do planeamento dentro do prazo previsto. Assim sendo o nosso projecto para a cadeira de Publicação Digital apresenta as seguintes componentes:

- **Interface de pesquisa à base de dados**

Esta interface possibilita ao utilizador pesquisar a Base de Dados de informação oticiosa impondo um conjunto de restrições.

- **Geração de Colecções de Tópicos**

Permite ao utilizador configurar a pesquisa que pretende fazer, de modo a que quando aceda ao URL lhe seja apresentado automaticamente um relatório com a informação desejada.

4. Arquitectura / Implementação

A Figura 3 apresenta o esquema global da arquitectura do Portal do NewsSearch.

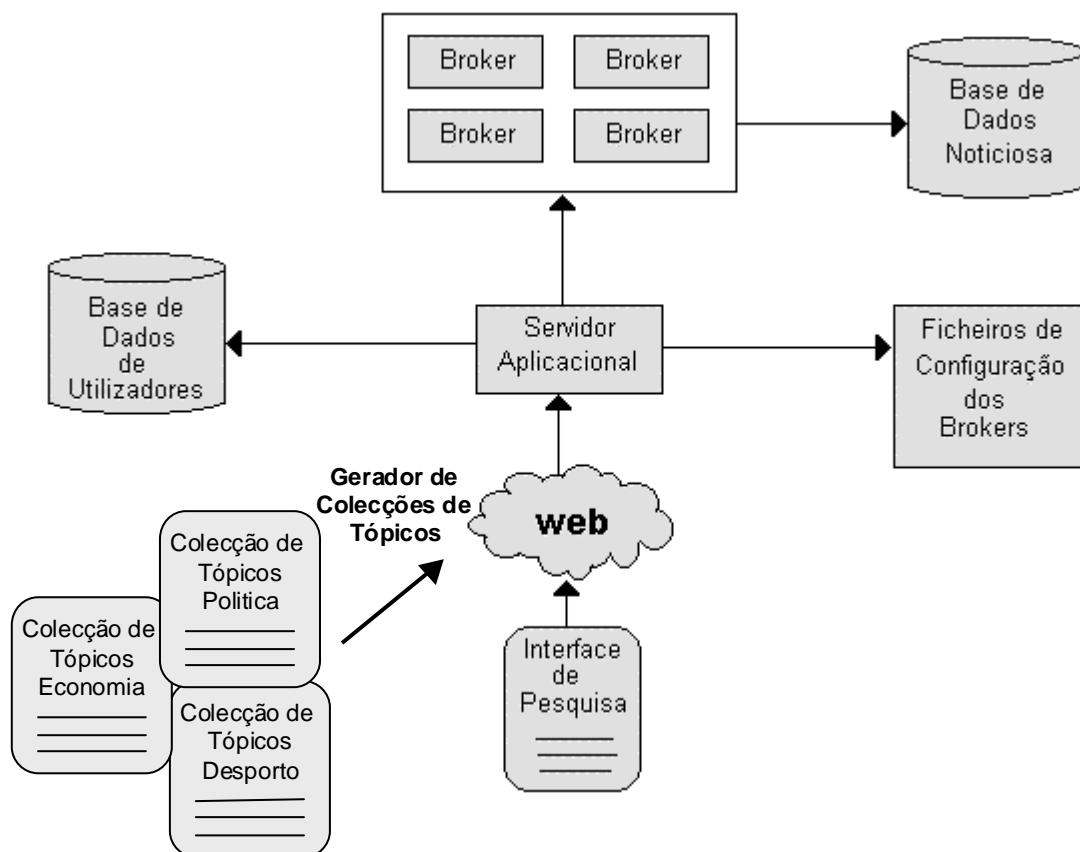


Figura 3. Arquitectura global do Portal do NewsSearch.

O Portal do NewsSearch oferece ao utilizador duas formas distintas de interacção com o sistema: a Interface de Pesquisa e o Gerador de Colecções de Tópicos.

Através da Interface de Pesquisa o utilizador envia as restrições de pesquisa escolhidas, o Servidor Aplicacional encarrega-se de criar as queries recorrendo aos ficheiros de configuração e envia-as aos Brokers. Os Brokers são processos que pesquisam a Base de dados Noticiosa, onde se encontra armazenada a informação acerca dos artigos recolhidos.

Caso o utilizador opte por recorrer ao Gerador de Colecções de Tópicos, este por sua vez recorre a uma Base de Dados de Utilizadores, onde se encontram armazenadas as definições de procura.

4.1 Interface de Pesquisa:

Esta componente recolhe informação acerca das restrições impostas pelo utilizador à pesquisa através de um formulário (ver Figura 4). Estas são recebidas e enviadas ao Servidor Apicacional que por sua vez consulta os ficheiros de configuração de cada Broker, usando esta informação e a recebida através da Interface para construir as *queries* a enviar a cada Broker.

Cada broker por sua vez usa o servidor de pesquisas do Glimpse para recolher e devolver as referências dos artigos que satisfazem a *query* ao Servidor Apicacional que os apresenta ao utilizador.

Pesquisa Avançada de Informação Noticiosa Notícias.PT

Termos a pesquisar
Ensino

Restrições

Categoria:

- Política
- Internacional
- Sociedade
- Regional
- Ciência
- Educação
- Desporto
- Computadores
- Economia
- Média
- Cultura

Publicação:

- Expresso
- Público
- Diário de Notícias
- Jornal de Notícias
- DN - Negócios
- O Jogo

Edição:
Última

Notícias.PT, FOUL - L18044
Conteúdo
Mon May 29 15:51:40 GMT+00:00 2000

Figura 4. Aspecto do formulário disponibilizado pelo Interface de Pesquisa.

As restrições à pesquisa podem ser especificadas através dos seguintes campos:

- **Publicação:** permite ao utilizador escolher as publicações sobre as quais a pesquisa deve incidir, tendo um leque de seis publicações noticiosas disponíveis na Web Portuguesa.
- **Categoria:** permite escolher as categorias de artigos que se pretende visualizar nos resultados, tendo um leque de 11 categorias cujas designações coincidem com as usadas para separar temas de interesse num jornal convencional.
- **Edição:** O utilizador poderá escolher receber artigos noticiosos cuja data de edição se encontre dentro de um dado intervalo temporal. São possíveis as opções: apenas a última edição, as edições dos últimos três dias, da última semana ou de todas as edições (sem restrição temporal).
- **Termos-chave:** O utilizador poderá procurar artigos que contenham um determinado termo, estes poderão conter operadores lógicos, por exemplo: Clinton AND Escândalo.

Os resultados da pesquisa são apresentados ao utilizador da seguinte forma: Em primeiro lugar é apresentado um resumo dos resultados obtidos para cada publicação (não visível na figura). De seguida são apresentados os resultados propriamente ditos organizados por publicação, sendo cada uma destas subdividida em categorias dentro das quais se encontram os cabeçalhos das notícias ordenadas por ordem de edição, sendo primeiro apresentados os artigos mais recentes.

A visualização dos resultados é paginada sendo dada a opção ao utilizador de visualizar mais artigos disponíveis segundo a restrição que os originou (caso existam), uma vez que são apresentados 10 artigos no máximo por categoria de modo a não tornar a recuperação de resultados demasiado lenta.

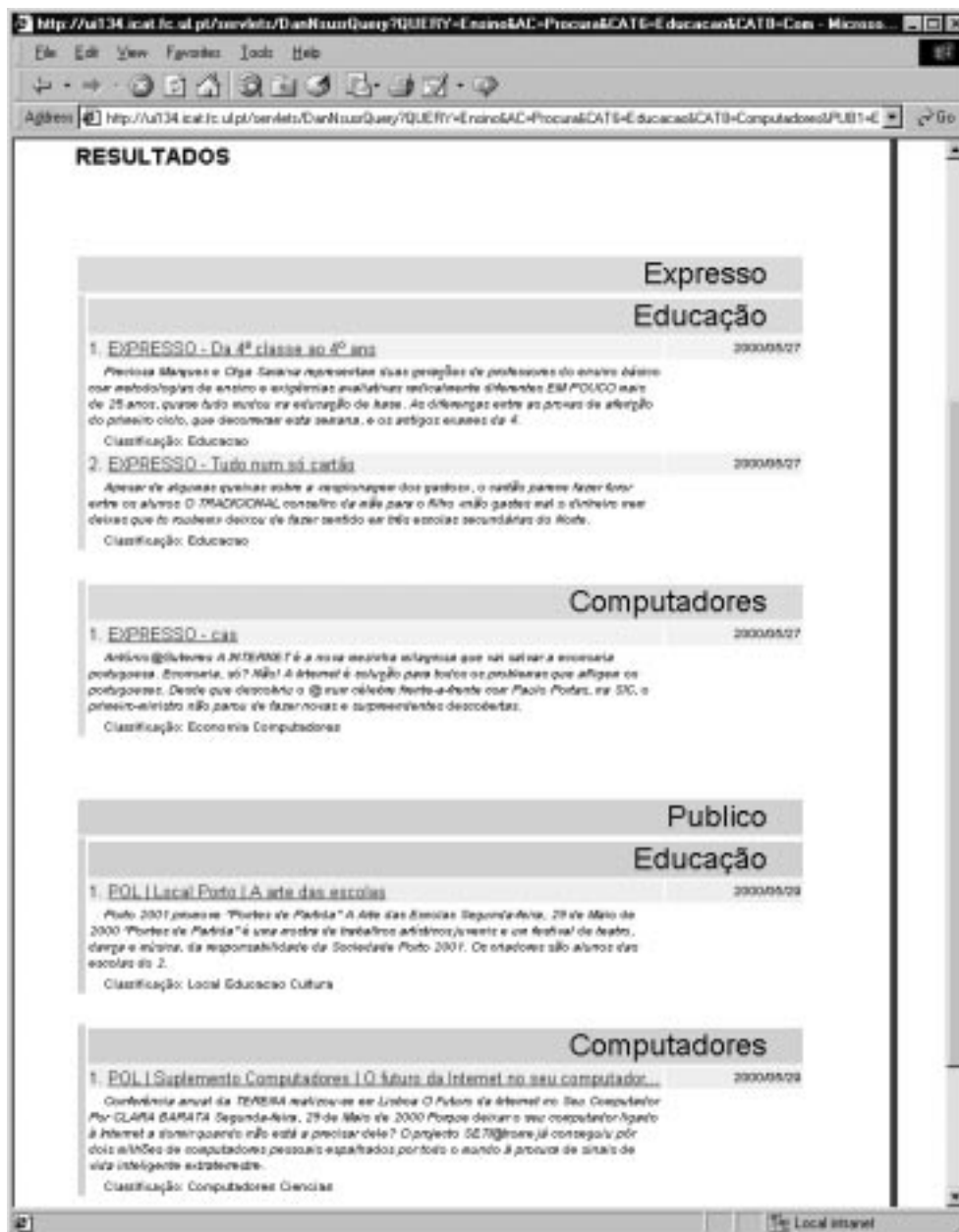


Figura 5. Aspecto dos resultado da pesquisa.

4.2 Gerador de Colecções de Tópicos

Esta componente permite a visualização imediata dos resultados de uma pesquisa na base de dados noticiosa pelo utilizador, para tal o Servidor Aplicacional comunica com o browser de modo a obter a cookie de identificação do utilizador:

- Caso esta exista, comunica com a base de dados de utilizadores de modo a receber as preferencias do utilizador, a informação obtida é então processada e a comunicação é redireccionada para um outro componente do Servidor Aplicacional responsável pela obtenção e visualização do relatório desejado.

- Caso o utilizador não esteja registado, o Servidor Aplicacional fornece-lhe um formulário de registo e configuração de restrições cuja a informação recolhida será guardada na base de dados de utilizadores e informa o browser de que o utilizador já se encontra registado, através do envio de uma cookie, mostrando em seguida ao utilizador os resultados obtidos a partir das restrições por ele configuradas.

Na Figura 5 apresentamos o modelo de dados que armazena as preferências dos utilizadores:

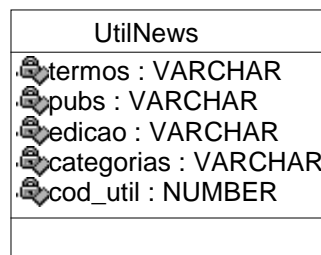


Figura 6. Modelo de dados para armazenamento dos dados dos utilizadores.

5. Resultados

O sistema encontra-se na fase de testes tendo já sido detectados alguns problemas e sido corrigidos erros entretanto encontrados. Encontra-se em funcionamento e disponível através da Internet o protótipo mais recente nas seguintes localizações:

- Interface de Pesquisa:
<http://ui134.icat.fc.ul.pt/servlets/DanNsusrSearch>
- Gerador de Colecções de Tópicos:
<http://ui134.icat.fc.ul.pt/servlets/newsCookie>

Foram ainda executados alguns testes de performance de forma a comparar estratégias na execução das pesquisas.

Numa primeira iteração do desenvolvimento deste projecto todas as *queries* eram enviadas para um único broker, tendo sido feitas as medições do tempo de resposta apresentados nos gráficos que se seguem.

Numa fase posterior julgou-se interessante estudar a possibilidade de explorar a capacidade de multiprocessamento da máquina onde está actualmente instalado o sistema NewsSearch (uma vez que dispõe de dois processadores).

Para este fim lançamos vários brokers, um por cada publicação, sendo cada *query* enviada para cada um deles, tendo sido obtidos os resultados para tempos de resposta ligeiramente inferiores.

Pela análise do gráfico podemos inferir que as nossas expectativas estavam correctas embora não tenham tido o impacto esperado, uma vez que apenas foi conseguido um melhoramento de cerca de 4,6% não sendo significativo para o aumento de memória necessário para acolher tantos brokers.

No entanto, este resultado é muito importante porque abre a perspectiva de melhorar o desempenho do sistema distribuindo-o, podendo vir a ser conseguidos resultados mais significativos.

No primeiro teste efectuámos um conjunto de medições que considerámos ser um padrão típico de utilização, ou seja, 1 restrição de termo chave, 3 de publicação, 3 de categoria e procura generalizada sem restrição de edição, tendo sido obtidos os tempos de resposta apresentados no Gráfico 1.

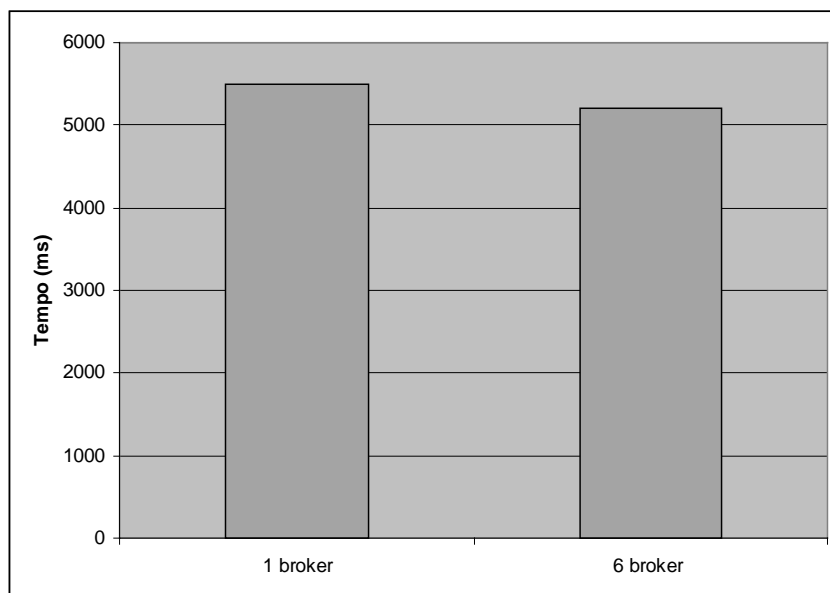


Gráfico 1. Média de tempo de resposta obtido partir de teste de utilização típica

Num segundo teste procurámos carregar ao máximo o sistema de busca fazendo uma pesquisa com 1 restrição por termo, todas as restrições por publicação (6) e por categoria (11) sendo esta busca efectuada sem restrição de edição.

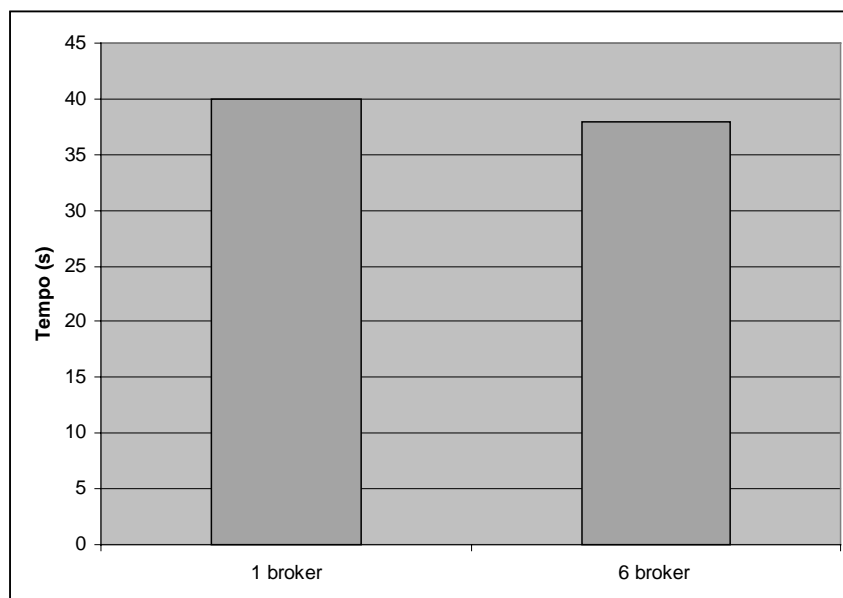


Gráfico 2. Média de tempo de resposta obtido partir de teste carga máxima

6. Conclusões e Trabalho Futuro

Neste projecto desenvolvemos uma Interface de Pesquisa e um Gerador Automático de Relatórios, tendo sido aberta a porta do sistema NewsSearch ao utilizador comum, que poderá utilizá-lo de forma proveitosa e simples, tornando-o quem sabe, no substituto dos “montes” de papel todas as manhãs acumulados na secretária, nos quais normalmente, não se encontra aquilo que se pretende.

Consideramos que foram obtidos resultados satisfatórios quer na facilidade com que alguns utilizadores menos experientes se familiarizaram com a utilização da Interface de Pesquisa, quer no desempenho obtido pelo sistema, uma vez que consideramos admissível o tempo médio de resposta de 5 segundos por uma busca típica.

No entanto, não aconselhamos que os utilizadores façam as pesquisas todas numa única interacção, uma vez que além de a resposta ser lenta devido ao tempo de processamento e principalmente à demora causada pela rede na transferência de resultados, a visualização dos resultados pode tornar-se confusa. Esta situação mereceria algum estudo de modo a estabelecer um padrão de utilização mais correcto e se necessário reduzir o numero de restrições simultâneas possíveis, mas estes aspectos teriam de ser estudados numa fase mais avançada do projecto na qual este estivesse mais difundido e estabilizado.

6.1 Trabalho Futuro

Como se verificou uma sobrecarga de trabalho imposta não só pelo desenvolvimento mas também por projectos de outras cadeiras, não foi possível a implementação do sistema de pesquisa de artigos similares.

Este módulo tem como objectivo preenchimento de relações de similaridade na base de dados artigos. Nesta tarefa além da consulta à base de dados pretende-se também o conclusão do mecanismo de detecção de similaridades do Serviço de Classificação, testando os artigos já armazenados na base de dados contra outros recebidos dos sites noticiosos. Pretende-se a realização de um teste de fiabilidade aos resultados obtidos e uma interface para acesso a esta informação.

O utilizador visualizará os artigos por ordem decrescente de relacionamento.

7. Referências

Glimpse Web site, <http://www.webglimpse.net>.

Korfhage R. Information Storage and Retrieval, Jonh Wiley & Sons, 1997.

Maria N. e Silva M. J. Theme-based Retrieval of Web News. Proceedings of the 13th Workshop on Web and Databases WebDB 2000.

Pressman R. Software Engineering. Fourth Edition. McGrawHill, 1997.