4
Ambiente de Experimentação

O ambiente de experimentação desenvolvido baseia-se em dois pontos principais, a saber:

- A implementação de uma ferramenta capaz de calcular as classificações das páginas da *WWW*, utilizando os algoritmos HITS e XHITS;
- A definição do *benchmark* utilizado para as comparações.

Com base nessa observação, apresentamos a seguir a ferramenta desenvolvida e o *benchmark* adotado.

4.1.
A Ferramenta

A ferramenta de experimentação consiste num programa desenvolvido na linguagem Java que permite efetuar consultas a máquinas de busca baseadas em texto, construir um sub-grafo direcionado da *WWW* e efetuar classificações de páginas do grafo a partir dos algoritmos *HITS* e *XHITS*.

Adicionalmente, foram implementados nesse ambiente alguns módulos de suporte para calibração dos parâmetros dos algoritmos, *HITS* e *XHITS*, e cálculo estatístico de acertos nas classificações das páginas.

Nos tópicos que se seguem, descrevemos o funcionamento de cada um desses módulos detalhadamente.

4.1.1.
SubGrafo da *WWW*

Este módulo tem por finalidade montar o grafo de páginas da *WWW* que servirá de base para as classificações dos algoritmos *HITS* e *XHITS*. Nesse grafo, as páginas são os vértices e os *hyperlinks* são as arestas. A criação do

grafo está baseada no modelo de subgrafo proposto por *Jon Kleingerg*. Assim, é preciso definir a palavra da consulta, o tamanho do núcleo inicial do grafo e o fator de corte para o número de páginas que apontam para as páginas do núcleo inicial.

Para a montagem do núcleo inicial, podemos definir, além do seu tamanho, se essas páginas são buscadas por um critério apenas textual, ou seja, por freqüência de repetição da palavra no conteúdo das páginas.

Outra opção existente é o acompanhamento e o registro do *log* do sistema enquanto esse efetua as buscas e extrai os *hyperlinks* das páginas, servindo para sanar quaisquer dúvidas sobre as páginas que foram integradas ao grafo.

Após a montagem do grafo, podemos salvá-lo integralmente, o que permite efetuar cálculos posteriores, sem a influência das mudanças da *WWW*. Então, quando não se desejar montar um novo grafo, basta carregar algum pré-existente, e este é automaticamente reconhecido pelos demais módulos.

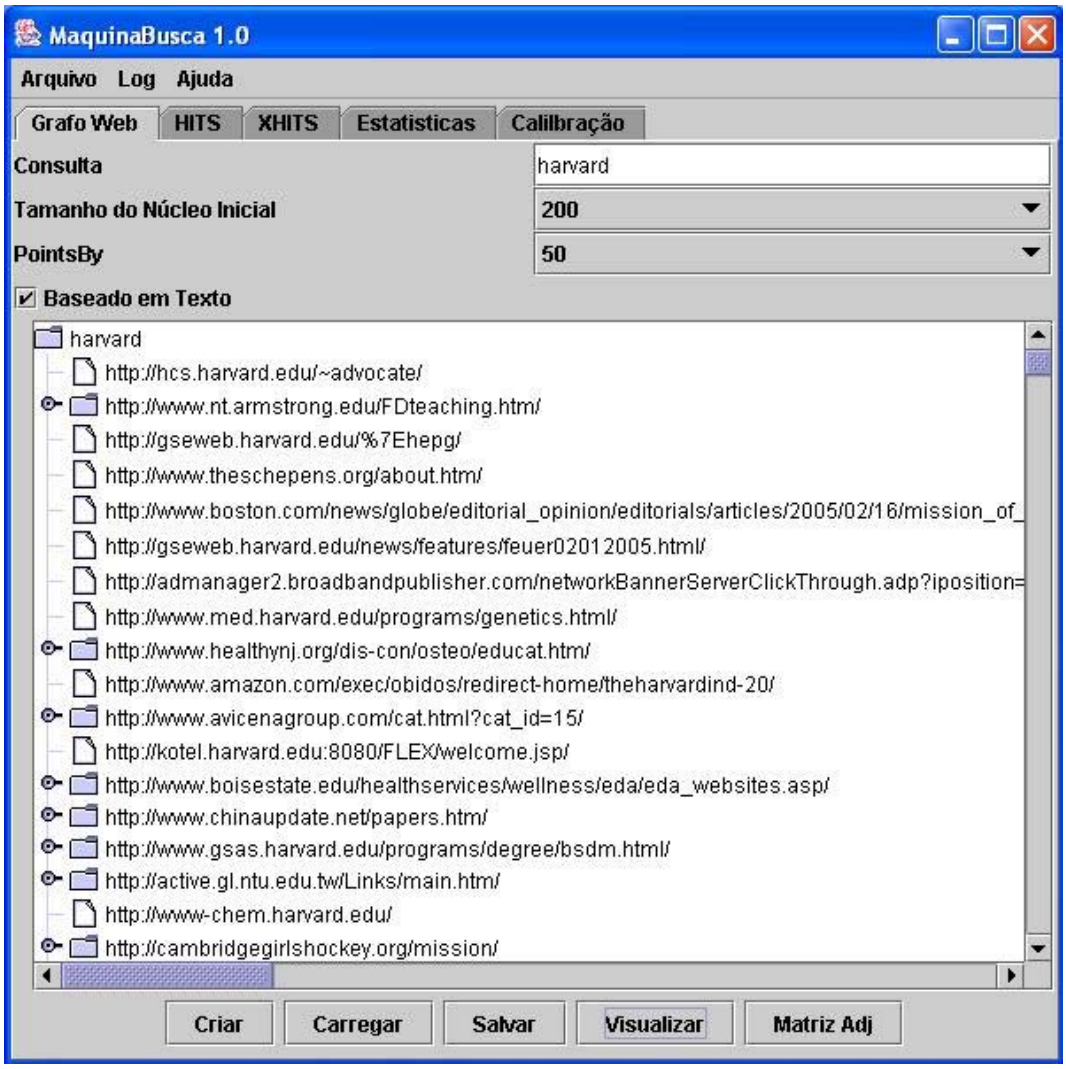

Figura 3: Módulo de construção do sub-grafo direcionado.

Na figura 3, vemos a montagem do grafo para a palavra *Harvard* e como podemos visualizá-lo através de uma árvore de diretórios. Cada arquivo e cada pasta representam uma página da *WWW*. Os arquivos internos a cada pasta significam que a página representada pela pasta possui um *hyperlink* para as páginas representadas pelos arquivos. Desta forma, podemos representar qualquer grafo resultante da consulta.

4.1.2. Algoritmo *HITS*

Neste módulo podemos efetuar a classificação das páginas que constituem o grafo gerado no módulo antes descrito, através do algoritmo *HITS*. Para tanto, algumas opções estão disponíveis para o usuário alterar de acordo com a sua

análise. O resultado da classificação está ordenado pelo valor da autoridade, conforme o modelo de *Hubs and Authorities*.

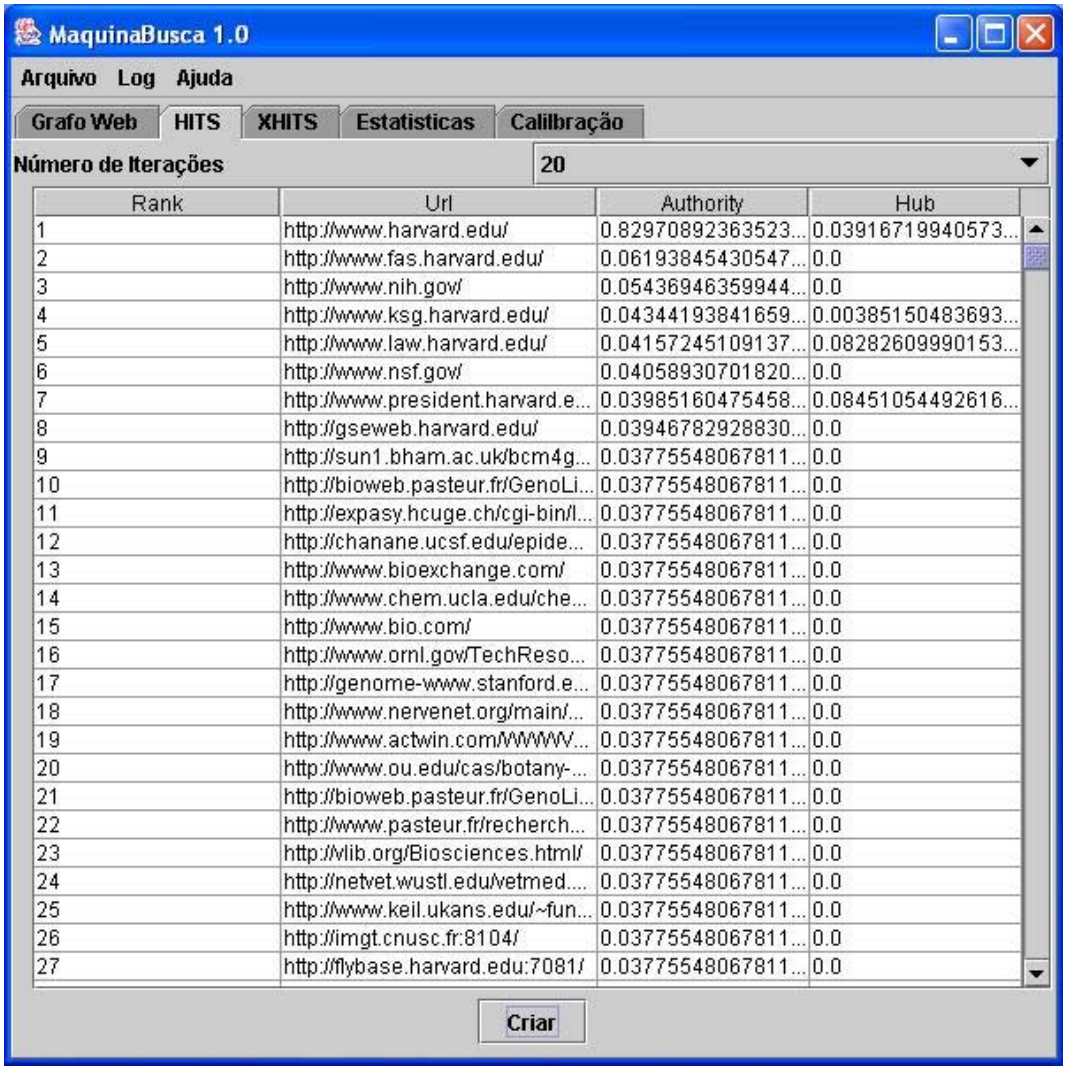

Figura 4: Classificação das páginas do grafo utilizando o algoritmo *HITS*.

Na figura 4, vemos a classificação das páginas do grafo ordenadas pelo valor da autoridade. Para gerar a classificação utilizando-se o HITS, basta definir qual grafo a ser utilizado e quantas iterações o algoritmo vai efetuar, pelos parâmetros G e k, respectivamente. Como a escolha do grafo já foi efetuada pelo módulo anterior, basta apenas definir o parâmetro k, escolhendo através da interface o número de iterações.

4.1.3. Algoritmo *XHITS*

Diferentemente do módulo do *HITS*, no qual era preciso apenas definir o número de iterações, o presente módulo possui alguns parâmetros adicionais em razão dos novos conceitos inseridos pelo *XHITS*. Esses novos parâmetros efetuam um ajuste fino nas relações entre autoridade, *hub*, portal e novidade.

Esse módulo nos permite definir um valor específico para cada um dos parâmetros de ajuste. Ou seja, é necessário definir o valor dos cinco parâmetros alfa (α), beta (β), gama (γ), fi (φ) e teta (θ) de acordo com as distribuições organizadas na matriz a seguir

$$M = \begin{bmatrix} 0 & A^T & \alpha A^T & \varphi A \\ A & 0 & \theta A^T & \beta A \\ \alpha A & \theta A & 0 & \gamma A \\ \varphi A^T & \beta A^T & \gamma A^T & 0 \end{bmatrix}$$

Podemos analisar os parâmetros, também, ao invés de olhar para a matriz, através dos seus significados semânticos:

- O parâmetro α relaciona autoridade e portal;
- O parâmetro β relaciona *hub* e novidade;
- O parâmetro θ relaciona *hub* e portal;
- O parâmetro φ relaciona autoridade e novidade;
- O parâmetro γ relaciona novidade e portal.

Para facilitar a interpretação dos parâmetros, tais relações semânticas foram descritas na interface, como mostrado na figura 5.

Definidos os valores dos parâmetros, basta definir o número de iterações, k, e efetuar a classificação.

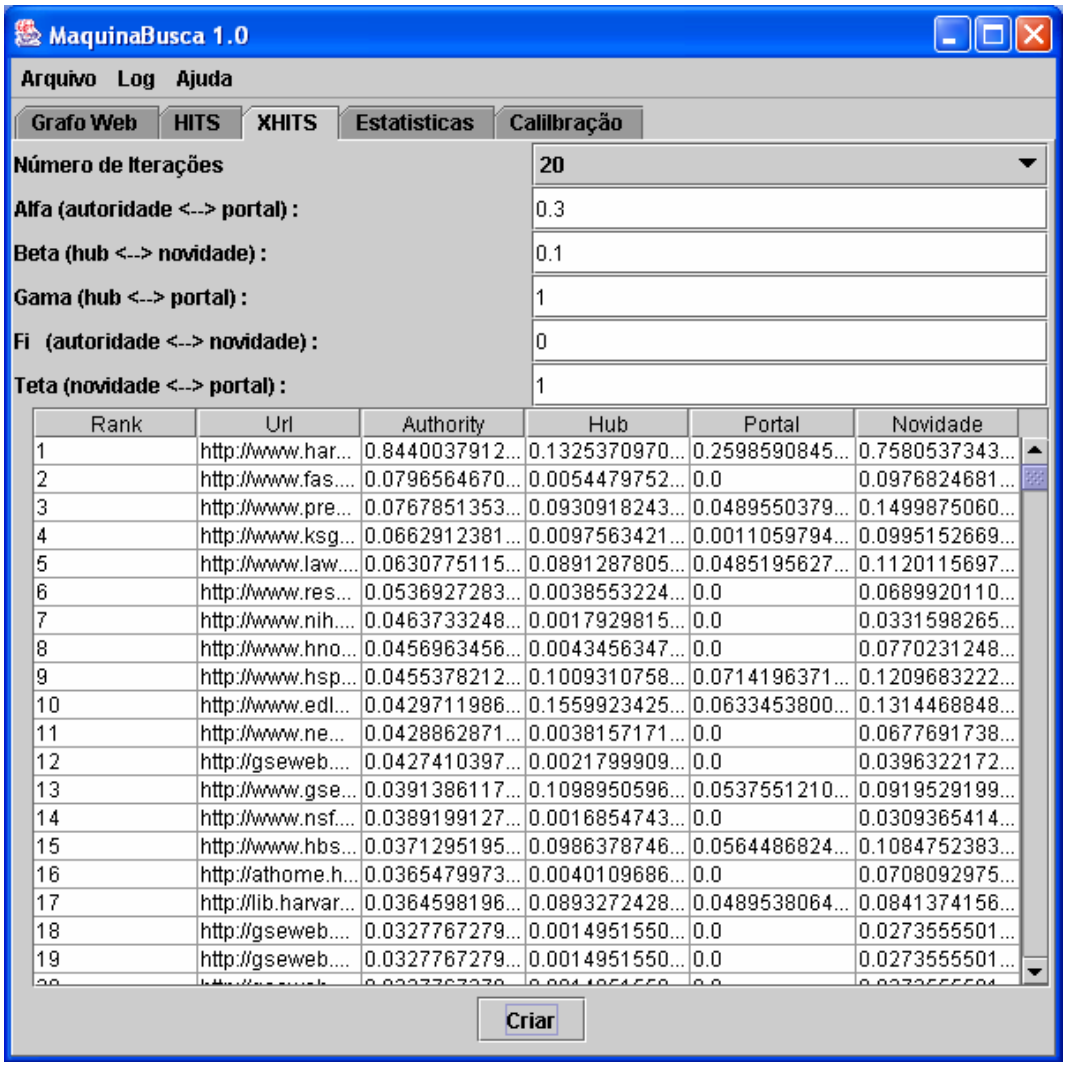

Figura 5: Classificação das páginas do grafo utilizando o algoritmo *XHITS*.

Na figura 5, vemos a classificação das páginas, do grafo antes definido, ordenadas pelo valor da autoridade.

4.1.4. Estatísticas e Gráficos

Nesse módulo, podemos efetuar os cálculos não só para o algoritmo *HITS*, como também, para o algoritmo *XHITS*. A partir de uma classificação fornecida previamente, podemos acompanhar a evolução dessas páginas conforme a variação dos parâmetros do algoritmo que estiver em análise por intermédio dos gráficos.

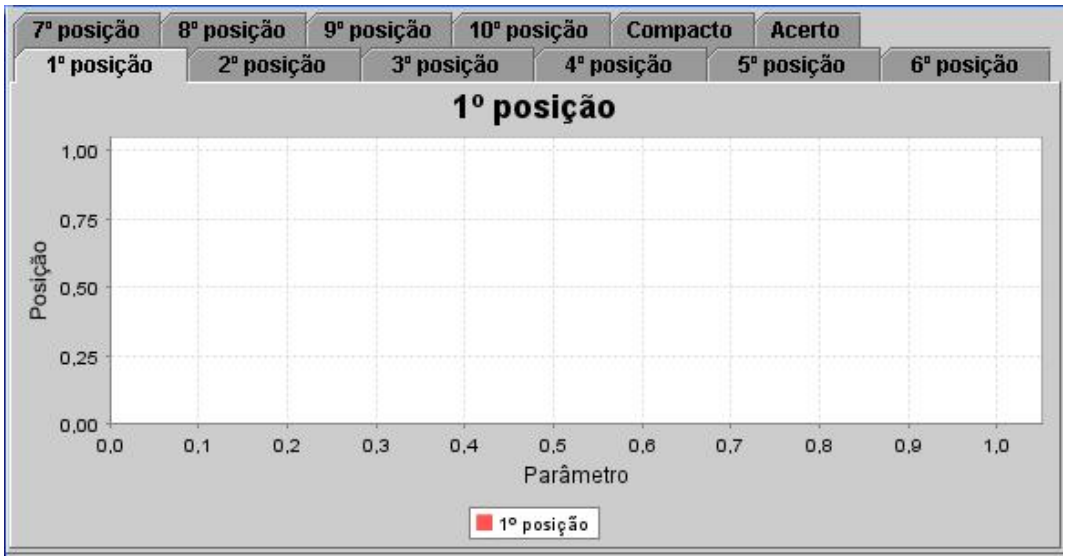

Figura 6: Gráficos comuns aos dois algoritmos.

Para ambos algoritmos, foi gerado um gráfico, figura 6, para cada uma das dez páginas fornecidas na classificação, um gráfico compacto que reúne a curva das dez páginas simultaneamente e um gráfico para cada intervalo de páginas que permite verificar a porcentagem de acertos por intervalo.

A seguir, veremos os sub-módulos para os algoritmos *HITS* e *XHITS* detalhadamente.

4.1.5. Algoritmo *HITS*

Nesse sub-módulo podemos fazer a variação do número de iterações do algoritmo *HITS*, definindo seu valor inicial, final e o passo utilizado para percorrer o intervalo. Além disso, temos que definir uma classificação de dez páginas para que se possa efetuar o acompanhamento através dos gráficos. Essa classificação, calculada externamente, permite avaliar o desempenho do HITS.

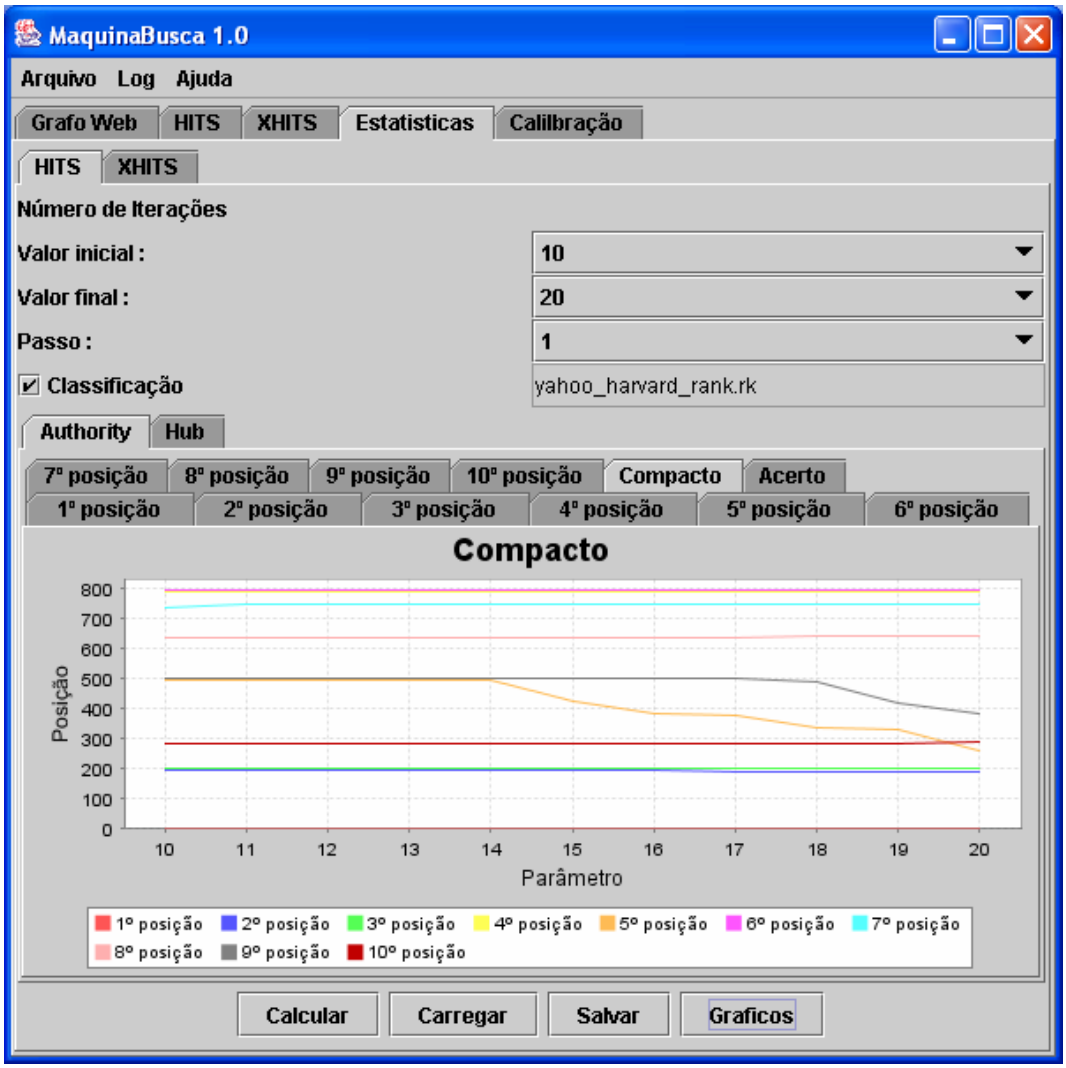

Figura 7: Acompanhamento da classificação pela variação das iterações.

Na figura 7, o valor inicial das iterações foi definido como dez e o valor final como vinte. A escolha de um para o passo gerou onze valores entre dez e vinte, como representado no eixo do parâmetro. O gráfico representado na figura 7 possui as curvas das dez páginas que estão sendo acompanhadas, cada uma de uma cor diferente com a sua respectiva legenda.

4.1.6. Algoritmo *XHITS*

Nesse sub-módulo podemos fazer a variação de um dos parâmetros do algoritmo *XHITS*, definindo seu valor inicial, final e o passo utilizado para percorrer o intervalo. Além disso, temos que definir uma classificação externa de dez páginas para que se possa efetuar o acompanhamento através dos gráficos.

Devemos, também, definir o número de iterações que o algoritmo vai efetuar para cada entrada diferente de parâmetros.

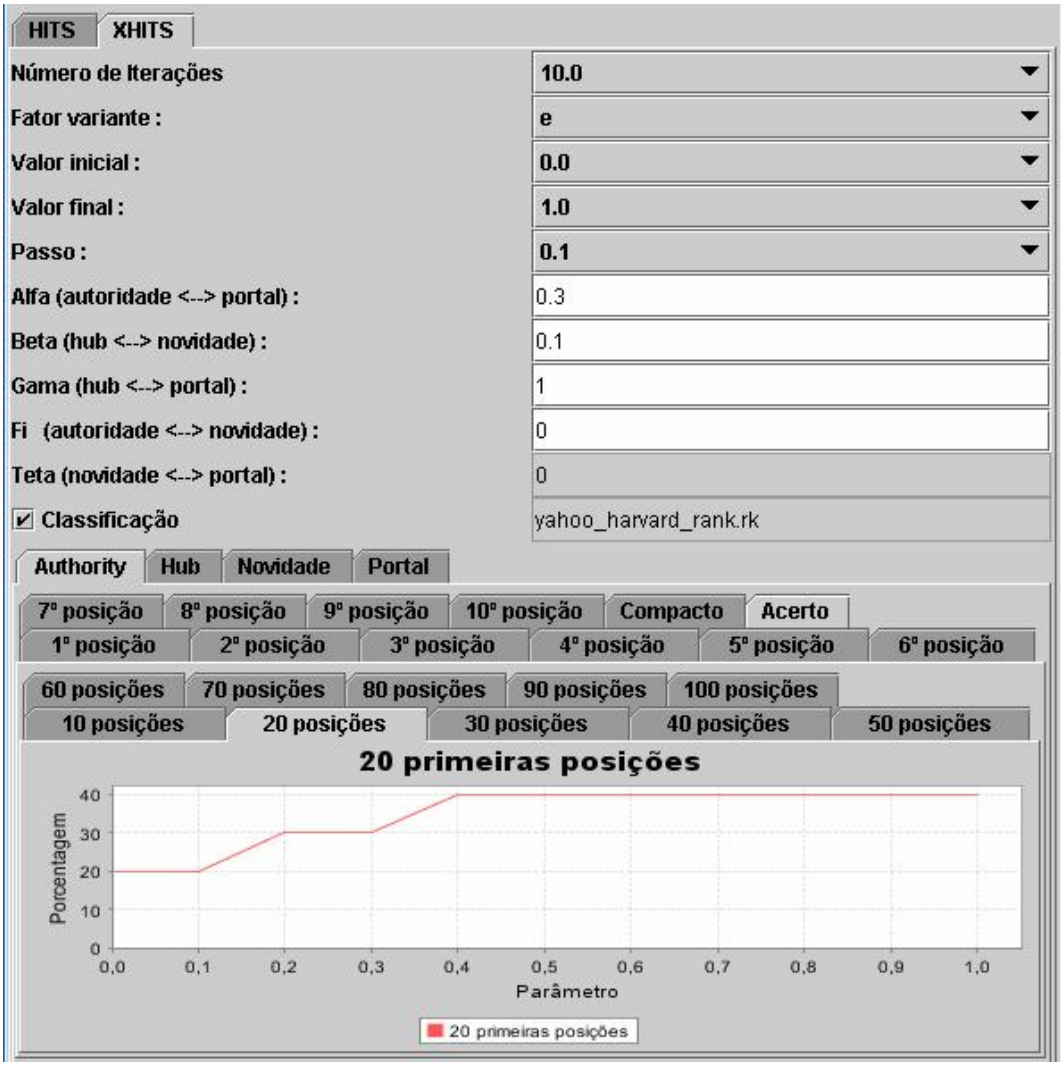

Figura 8: Acompanhamento da classificação pela variação do parâmetro teta.

Na figura 8, o valor das iterações foi fixado em 20, o valor inicial do parâmetro θ foi definido como zero, o valor final como um e o seu passo como um décimo, gerando onze valores entre zero e um, representado no eixo do parâmetro. O gráfico representado na figura 8 indica a porcentagem de acertos de acordo com o aumento no valor do parâmetro teta. Os demais parâmetros também foram definidos e ficaram invariantes ao longo do cálculo.

4.1.7. Calibração

Como visto nos módulos anteriores, o algoritmo *XHITS* possui cinco parâmetros que precisam ser ajustados (figura 9), ou seja, a partir de uma classificação fornecida deseja-se encontrar o conjunto de parâmetros que maximiza a porcentagem de acertos (figura 9), no menor intervalo de páginas possível. Por esta razão, esse módulo foi desenvolvido para varrer o espaço de soluções buscando esses parâmetros maximizadores.

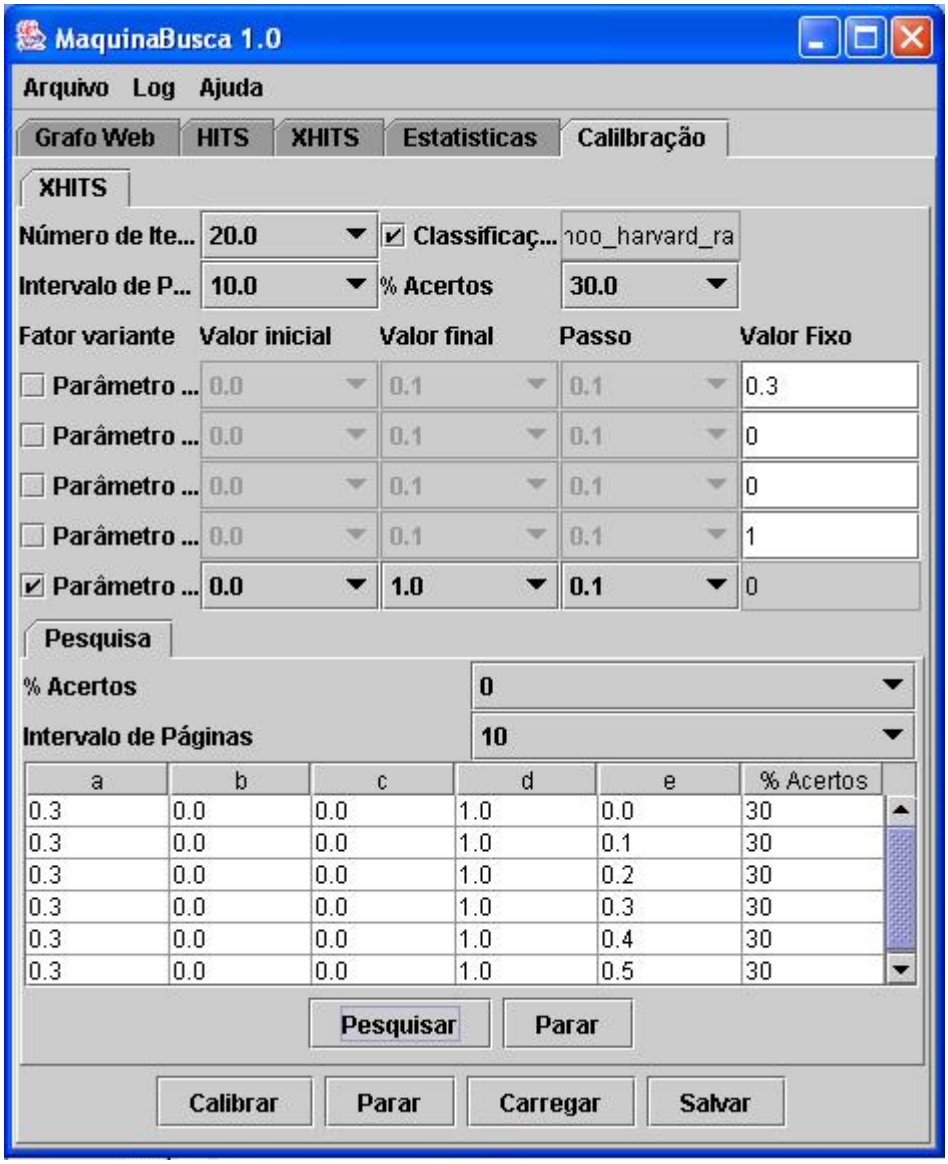

Figura 9: Calibração dos parâmetros do *XHITS*.

O tamanho do espaço de soluções é da mesma ordem que o resultado da multiplicação dos números de valores para cada um dos parâmetros. Por

exemplo, se cada um tiver dez valores diferentes, a ordem será de cem mil valores e se para cada iteração for gasto meio segundo, o tempo total, de computação, será, aproximadamente, de 14 horas.

Além do tempo, existe a questão do tamanho da memória para manipular tais valores. Para isso, implementou-se um filtro que determina se a solução parcial calculada deve ou não ser guardada. O filtro leva em consideração a porcentagem mínima de acerto num determinado intervalo. Por exemplo, podemos guardar apenas a conjugação de parâmetros que tenha pelo menos 30% de acerto nas dez primeiras páginas.

Sendo assim, a critério de quem estiver analisando os dados, podem-se definir valores variantes ou fixos para cada um dos parâmetros e, também, um limite inferior de acertos para os resultados.

Adicionalmente, para melhor procura da solução, foi implementada uma interface de pesquisa que varre o espaço de soluções calculado e retorna numa tabela as soluções que satisfazem os critérios estabelecidos nos campos da pesquisa. Nessa interface existem dois campos: porcentagem de acertos e intervalo de páginas. O primeiro define a porcentagem mínima de acertos que a solução deve satisfazer. O segundo define em qual intervalo de páginas essa porcentagem deve ser satisfeita, por exemplo, trinta por cento de acerto nas dez primeiras páginas.

4.2. Benchmark

Nos tópicos que se seguem, descrevemos as consultas utilizadas nessa dissertação, os especialistas utilizados para as classificações de comparação e o critério de relevância adotado para avaliar os resultados.

4.2.1. Consultas

A correta escolha da base de teste é de extrema importância, uma vez que o dinamismo da *WWW* interfere diretamente nas bases de dados. Tal dinamismo decorre das atualizações e das novas inserções de páginas *HTML* na *WWW*.

Como essas bases modificam-se com o passar do tempo [23,24,25,26,27], torna-se necessária a sua conservação.

Desta forma, para cada consulta é gerado um sub-grafo da *WWW* que guarda em si a estrutura de *hyperlinks* no momento da consulta. Sendo assim, pode-se salvar esse sub-grafo, congelando uma pequena parte da *WWW*, para posteriores consultas e cálculos.

Em relação aos tópicos de busca, foram extraídos seis diferentes tópicos baseados na literatura:

- *Abortion;*
- *Artificial Intelligence;*
- *Harvard*
- *Jaguar;*
- *Kyoto University;*
- *Olympic.*

Porém, se torna mais difícil avaliar o conteúdo das classificações com este tipo de tópico, pois pertencem a domínios estrangeiros. Por tal motivo, optamos pela utilização de tópicos em português, que teriam a mesma amplitude dos tópicos em inglês, ressalvado o tópico *Harvard*, para fins de comparação.

Desta forma, para esse trabalho foram definidos os tópicos de busca a seguir:

- Fome;
- *Harvard;*
- Rio de Janeiro;
- Puc-rio;
- Lula;
- Pelé;
- Marinha;
- Exército.

Tais tópicos foram utilizados posteriormente como conjunto de treinamento do modelo estendido, XHITS e como forma de validação e avaliação do treinamento foram definidas outras vinte e três consultas, quais sejam:

- "tribunal de justiça";
- "referendo do desarmamento";
- "balão mágico";
- "criança esperança";
- carambola;
- Petrópolis;
- mamífero;
- mensalão;
- xuxa;
- java;
- telemar;
- vivo;
- claro;
- gato;
- cachorro;
- agricultura;
- Brasil;
- gurgel;
- jogos;
- itamaraty;
- celular;
- catolicismo;
- ametista;

A seguir as considerações feitas para a escolha dos especialistas e a importância desses nesse trabalho.

4.2.2.
Avaliação dos Especialistas

A fase seguinte engendra um problema delicado, pois o julgamento da relevância das páginas classificadas pelos métodos implementados é subjetivo. Assim, algumas classificações devem ser consideradas como referência.

O custo associado aos especialistas humanos para obtenção de julgamentos de qualidade das respostas é extremamente elevado. Portanto, foram escolhidos dois especialistas artificiais para efetuar as classificações de referência:

- *Google*;
- *Yahoo*.

A partir das classificações fornecidas por estas duas máquinas de busca, foram efetuadas comparações de qualidade e, dentro do possível, análises humanas do conteúdo das páginas melhores classificadas.

4.2.3.
Relevância

As classificações geradas pelos algoritmos foram divididas em intervalos de dez páginas e, para efeitos de visualização, são mostrados os dez primeiros intervalos. Entretanto, para efeito de análise dos acertos, foram considerados, para as resultados finais, apenas os cinco primeiros intervalos, ou seja, as cinqüenta primeiras páginas classificadas. Tal escolha decorre da busca usual na *WWW*, em que o usuário comum, em geral, não averigua além dos vinte ou trinta primeiros resultados retornados na busca[20,21,22].

Nesse ponto, carregamos os dez primeiros resultados das máquinas do *Google* e do *Yahoo* e identificamos a incidência dessas páginas nos dez primeiros intervalos, ou seja, nas cem primeiras páginas classificadas pelo ambiente. Posteriormente, o índice de acerto por intervalo é calculado e utilizado na análise dos resultados.

Na próxima seção encontramos as análises e os resultados das buscas efetuadas a partir dos tópicos antes citados.