5
Experimentos

Passamos, nesse capítulo, à análise dos resultados experimentais e à sua discussão. Para tanto, como mencionado anteriormente, foram definidos os seguintes tópicos de busca para a fase de treinamento e análise:

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

Os grafos para cada tópico foram gerados, de acordo com os valores fornecidos nos experimentos de *Jon Kleinberg* [1,2], com os seguintes parâmetros:

- Núcleo inicial com 200 páginas;
- O número de páginas que apontam para as páginas do núcleo é igual a 50.

Em seguida, para cada tópico foram coletadas as classificações das dez primeiras páginas dos especialistas, *Google* (http://www.google.com.br) e *Yahoo* (http://www.yahoo.com.br), de forma a efetuarmos as comparações e análises necessárias.

Em prosseguimento, uma calibração foi efetuada para se encontrar valores adequados para os parâmetros que resultassem no maior número de acertos nas dez, vinte e cinqüenta primeiras páginas.

Diante dos resultados da calibração foram extraídos os valores para cada parâmetro de redução. Os valores de α, β, θ, φ, γ, com maior freqüência, que maximizam o índice de acertos nas dez primeiras páginas, foram obtidos através de um histograma. Desse conjunto resultante de valores, foram extraídas duas alternativas para os parâmetros, denominadas Solução Histograma Média e Solução Histograma Modal.

As duas soluções definidas foram aplicadas a um conjunto de vinte e três consultas e o desempenho avaliado em relação ao especialista *Google*. A escolha desse especialista foi somente pela facilidade de automação do processo, pois o *Google* possui uma interface de consulta disponível em *Java*, possibilitando a busca do ranking de comparação pela ferramenta desenvolvida.

A seguir apresentamos um detalhamento das soluções antes mencionadas, em seguida uma análise agrupada dos resultados e, posteriormente, uma análise mais detalhada de dois dos oito tópicos antes definidos.

5.1. Soluções por Histograma Média e Modal

Inicialmente, o espaço de soluções total para a calibração foi discretizado em valores entre -1 e 1 para cada parâmetro, com passo de 0,5. Com essa discretização, e pelo fato de variarmos cinco diferentes parâmetros, reduzimos o espaço de soluções para 5^5 (15625) soluções.

Sendo assim, a partir das oito consultas para o treinamento, para cada uma delas, foram gerados dois conjuntos de valores que maximizam o índice de acerto, superiores a vinte por cento, para as dez primeiras páginas em comparação com as classificações do *Google* e do *Yahoo*. Dessa forma, ao final temos 16 conjuntos de valores possíveis para os parâmetros do *XHITS*. Em termos práticos, cada conjunto gerou em média duzentos e cinqüenta soluções diferentes.

Como próximo passo, os valores desses dezesseis conjuntos foram inseridos em um histograma para se definir as soluções que apresentavam maior freqüência em todos os conjuntos. Dessa forma, ao percorrer os conjuntos, um

ponto foi adicionado para cada solução encontrada. Como temos dezesseis conjuntos, que individualmente não possuem respostas repetidas, a pontuação máxima que uma solução pode ter é dezesseis. Em nosso experimento, encontramos sete soluções que possuem freqüência igual a doze (tabela 1).

Solução	α	β	θ	Φ	γ
Histograma	-0,5	-1,0	1,0	0,0	-0,5
	-0,5	-1,0	1,0	1,0	0,5
	-0,5	-0,5	0,5	1,0	0,0
	-0,5	-0,5	0,5	1,0	0,5
	-0,5	-0,5	1,0	1,0	0,5
	-0,5	0,0	1,0	1,0	0,5
	0,0	-1,0	0,5	1,0	1,0
Histograma Média	-0,4285	-0,6428	0,7857	0,8571	0,3571
Histograma Modal	-0,5	-0,5	1,0	1,0	0,5

Tabela 1: Soluções por Histograma.

Após coletar as soluções de maior freqüência, as reduzimos à uma única solução de duas maneiras diferentes. A primeira, Solução por Histograma Média consiste em calcular a média dos valores encontrados por parâmetro. A segunda, Solução por Histograma Modal consiste em verificar, por parâmetro, qual valor ocorreu com maior freqüência, ou seja, a moda. Em ambas as soluções, ao final obtemos apenas um único valor para cada solução. Tanto a solução média, quanto a modal utilizada no experimento, encontram-se na tabela 1 acima.

A seguir vemos a discussão do desempenho destas soluções em comparação com o desempenho do *HITS*.

5.2. Resultados Experimentais

Para uma análise mais apurada dos resultados obtidos para as consultas de treinamento, adotamos duas medidas: relevância e concordância.

A relevância define a porcentagem de páginas retornadas com sucesso em relação à classificação de cada especialista num determinado intervalo de páginas. Formalmente, temos:

$$RL_i = \left[\frac{e_i}{n} \right] \times 100 , \qquad i = 10,20,...100$$

onde e_i é o número de páginas retornadas com sucesso, n é número de páginas do especialista e i é o intervalo de páginas.

A concordância, por sua vez, avalia o grau de identidade das avaliações dos especialistas. Para isso, como a quantidade de páginas que cada especialista avaliou é igual, basta calcular quantas páginas os especialistas possuem em comum e dividir pelo número de páginas que cada um avaliou. Assim, temos:

$$C = \left[\frac{pc}{to} \right] \times 100$$

onde pc é o número de páginas em comum dos dois especialistas e to é número de páginas avaliada por especialista.

Com estas medidas podemos, então, avaliar com maior precisão os resultados retornados pelo *XHITS* nas consultas de treinamento. Na tabela 2 e 3, podemos verificar a porcentagem de acertos correspondentes aos dois algoritmos em relação aos dois especialistas escolhidos. Então, para cada intervalo de páginas, possuímos a relevância do algoritmo *HITS* e do algoritmo *XHITS* por especialista.

		Yahoo					
		RL_{10}		**RL_{20}**		**RL_{50}**	
Tópicos	**Concordância**	**H**	**X**	**H**	**X**	**H**	**X**
Fome	70%	0%	60%	0%	60%	0%	70%
Pelé	70%	0%	50%	0%	50%	0%	90%
Puc-rio	70%	20%	40%	20%	50%	30%	80%
Rio de Janeiro	20%	10%	30%	10%	40%	10%	40%
Harvard	40%	10%	30%	10%	50%	10%	80%
Lula	60%	0%	20%	0%	40%	0%	40%
Marinha	50%	20%	60%	30%	70%	30%	100%
Exército	40%	20%	50%	20%	70%	20%	70%

Tabela 2: Comparações dos diversos resultados obtidos com o *Yahoo*.

		Google					
		RL_{10}		**RL_{20}**		**RL_{50}**	
Tópicos	**Concordância**	**H**	**X**	**H**	**X**	**H**	**X**
Fome	70%	0%	30%	0%	30%	0%	50%

Pelé	70%	0%	50%	0%	60%	0%	80%
Puc-rio	70%	20%	30%	20%	40%	20%	60%
Rio de Janeiro	20%	0%	40%	0%	40%	0%	40%
Harvard	40%	20%	10%	20%	10%	20%	30%
Lula	60%	0%	20%	0%	20%	0%	20%
Marinha	50%	30%	40%	30%	60%	40%	100%
Exército	40%	30%	30%	30%	30%	30%	30%

Tabela 3: Comparações dos diversos resultados obtidos com o *Google*.

Podemos observar nas tabelas 2 e 3 que, em todos os intervalos de páginas, o algoritmo *HITS* não conseguiu ultrapassar a margem de trinta por cento de relevância. Por outro lado, o algoritmo *XHITS*, no intervalo das dez primeiras páginas por exemplo, já possui no tópico fome, sessenta por cento de acerto, chegando a noventa por cento no intervalo das cinqüenta primeiras no tópico Pelé.

Nos tópicos fome, Pelé e Lula, o algoritmo *HITS* não conseguiu se aproximar em nada das classificações dos especialistas, resultando num índice de zero por cento em todos os intervalos, enquanto o *XHITS* conseguiu nas dez primeiras páginas, índices de sessenta, cinqüenta e vinte por cento, respectivamente.

Nos tópicos PUC-Rio, Rio de Janeiro e *Harvard*, o algoritmo *HITS* atingiu nas dez primeiras páginas, vinte, dez e dez por cento, respectivamente, contra quarenta, trinta e trinta por cento alcançados pelo *XHITS*.

No tópico marinha o XHITS consegue retornar todas as páginas dos especialistas, obtendo um índice de cem por cento de acerto para o *Google* e o *Yahoo*.

Mesmo com um baixo grau de concordância entre os especialistas no tópico Rio de Janeiro, o *XHITS* conseguiu equilibrar os seus acertos entre os dois especialistas. Recuperou trinta por cento nas dez primeiras páginas em relação ao *Yahoo* e quarenta no mesmo intervalo em relação ao *Google*.

Cabe ressaltar que somente em um dos intervalos de páginas, o algoritmo *HITS* esteve à frente do algoritmo *XHITS* em porcentagem de relevância e que em quase todos os demais intervalos, o XHITS manteve o dobro de porcentagem.

A seguir as dez primeiras páginas recuperadas pelo algoritmo *XHITS*, para os tópicos Lula, Pelé e Rio de Janeiro:

Lula		
Pos	**URL**	**Descrição**
1	http://www.cidob.org/bios/castellano/lideres/s-001.htm/	Centro de investigación, docencia, documentación y divulgación de Relaciones Internacionales y Desarrollo
2	http://www.lula.ca/	Lula Lounge
3	http://www.luladance.theblog.com.br/dancapoder3.swf/	Sátira ao presidente Lula
4	http://www.kirjasto.sci.fi/carsonmc.htm/	Lula Carson Smith
5	http://www.lascivalula.com.br/	LASCIVA LULA
6	http://www.lulacafe.com/	Lula Café
7	http://www.mapquest.com/maps/map.adp?city=Chicago&state=IL&zipcode=60647&address=2537%20N%2E%20Kedzie&country=us&zoom=8/	Mapa de localização do Lula Café
8	http://www.amazon.com/exec/obidos/redirect-home/authorscalend-20/	
9	http://www.yorku.ca/cerlac/recent03-04.html/	Globalization and Social Movements: A Brazilian Perspective
10	http://beerdrinkers.co.uk/	Portal

Tabela 4: Os dez primeiros resultados do *XHITS* para o tópico Lula.

Pelé		
Pos	**URL**	**Descrição**
1	http://www.explore-biography.com/sports_figures/P/Pel%E9.html/	Dictionary of Famous People
2	http://www.360soccer.com/pele/	360 soccer
3	http://www.latinosportslegends.com/Pele_bio.htm/	Latino Legends in Sports
4	http://www.encyclopedia.com/html/P/Pele2.asp/	Enciclopédia
5	http://www.time.com/time/time100/heroes/profile/pele01.html/	Time on line Edition
6	http://www.who2.com/pele.html/	guide to facts about famous people
7	http://noticias.uol.com.br/pelenet/	Página do Pelé na Uol
8	http://www.360soccer.com/pele/pelebio.html/	360 soccer
9	http://www.fifa.com/	Fifa
10	http://www.us-soccer.com/	Site oficial de Futebol dos EUA

Tabela 5: Os dez primeiros resultados do *XHITS* para o tópico Pelé.

Rio de Janeiro		
Pos	**URL**	**Descrição**
1	http://ipanema.com/	Informações turísticas sobre o Rio de Janeiro
2	http://www.riotransito.com.br/	Informações turísticas sobre o Rio de Janeiro
3	http://www.uerj.br/	Universidade do Estado do Rio de Janeiro
4	http://www.ufrj.br/	Universidade Federal do Rio de Janeiro
5	http://www.governo.rj.gov.br/	Página oficial do Estado do Rio de Janeiro
6	http://www.alerj.rj.gov.br/	Página oficial da Assembléia Legislativa do Estado do Rio de Janeiro
7	http://www.rio.rj.gov.br/	Página oficial da Prefeitura da cidade do Rio de Janeiro
8	http://www.via-rio.com.br/	Guia de eventos da cidade do Rio de Janeiro

9	http://www.mamrio.com.br/	Museu de Arte Moderna do Rio de Janeiro
10	http://www.jbrj.gov.br/	Instituto de Pesquisas Jardim Botânico do Rio de Janeiro

Tabela 6: Os dez primeiros resultados do *XHITS* para o tópico Rio de Janeiro.

Podemos verificar pelas tabelas 4, 5 e 6 a qualidade das páginas retornadas pelo algoritmo *XHITS* juntamente com a descrição destas. Nos resultados do tópico Lula, o *XHITS* conseguiu recuperar páginas que versam sobre o presidente Lula, pessoas famosas com nome e sobrenome Lula e o estabelecimento internacionalmente conhecido Café Lula.

Em relação ao tópico Pelé, todas as páginas que o *XHITS* recuperou versam sobre o Pelé, jogador de futebol, ou estão relacionadas com futebol no caso da FIFA. Já os resultados para o tópico Rio de Janeiro, três páginas versam sobre informações turísticas da cidade do Rio de Janeiro, duas são as principais universidades públicas do estado do Rio e as demais são páginas oficiais da Prefeitura do Rio de Janeiro, do estado do Rio de Janeiro, do Instituto de Pesquisas do Jardim Botânico e do Museu de Arte Moderna, aparentando um excelente resultado.

Em seguida, após as avaliações sobre o conjunto utilizado para o treinamento, aplicamos as Soluções por Histograma Média e Modal, e o *HITS* a um conjunto de vinte e três consultas aleatoriamente definidas. O desempenho dessas soluções foi baseado no especialista *Google*, pelos mesmos motivos já apresentados.

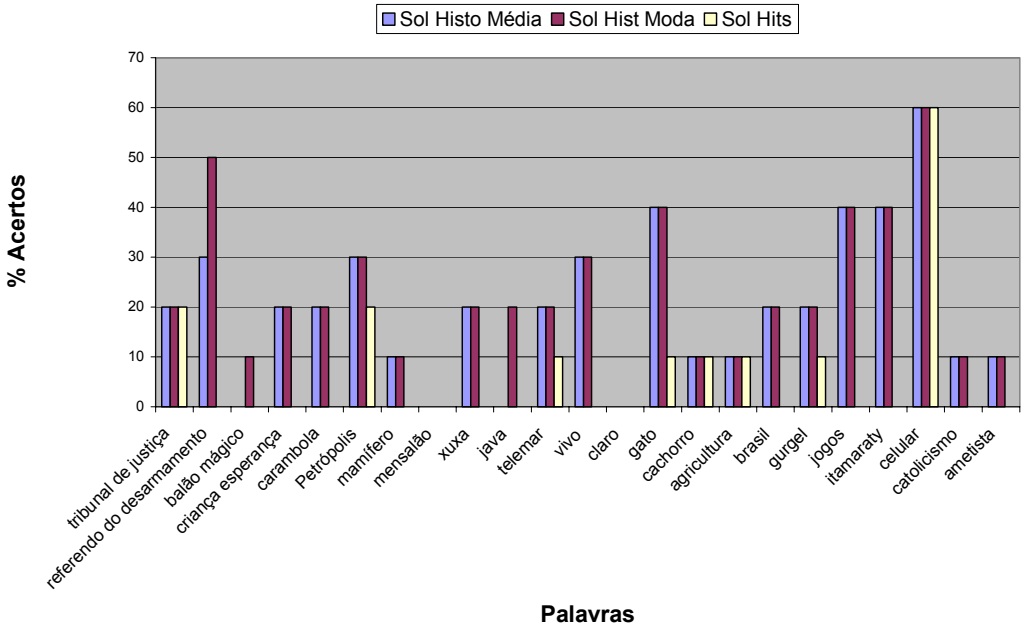

Figura 10: Gráfico de acertos por consulta nas dez primeiras páginas.

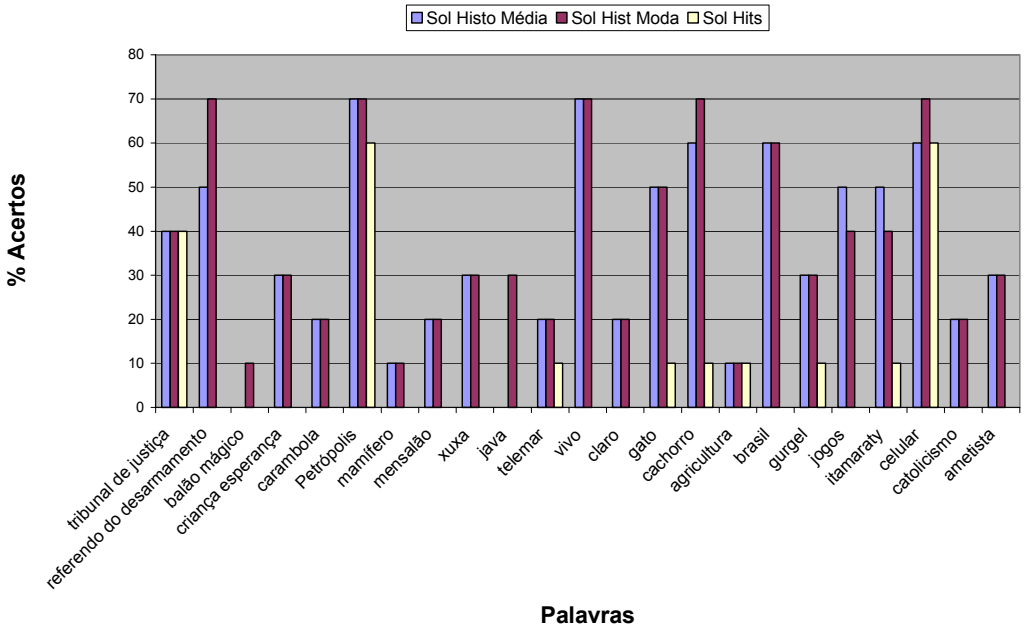

Figura 11: Gráfico de acertos por consulta nas vinte primeiras páginas.

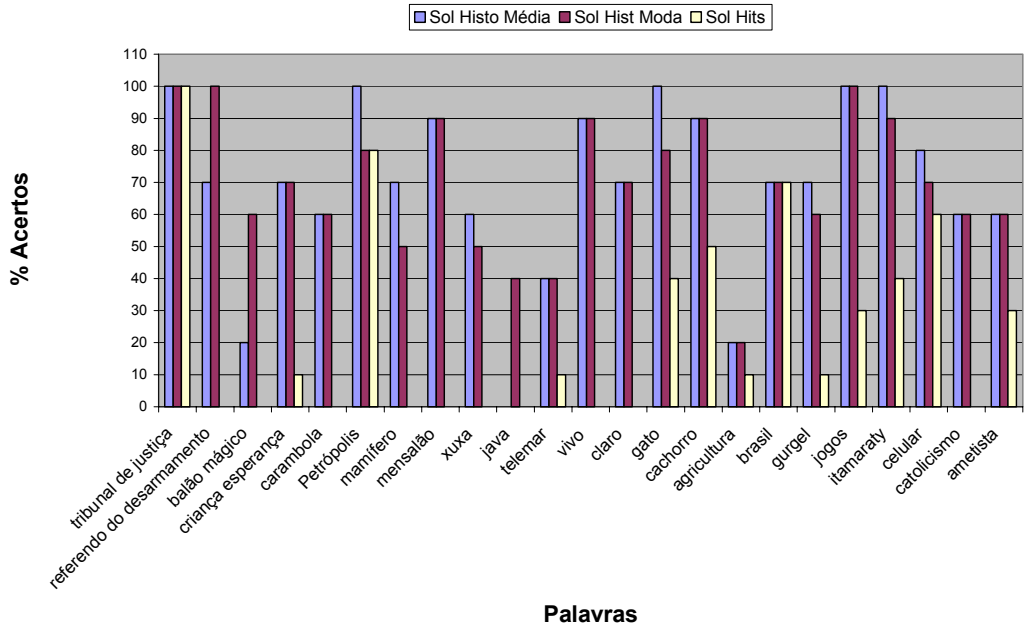

Figura 12: Gráfico de acertos por consulta nas cinqüenta primeiras páginas.

Nas figuras 10, 11 e 12 podemos verificar que o desempenho das soluções por histograma média e modal é superior ou, em alguns casos, igual ao do *HITS*. Na figura 10, vemos que as soluções empataram os seus desempenhos em seis consultas, sendo que em duas das seis nenhum acerto foi verificado. Aumentando o intervalo para as vinte primeiras, figura 11, vemos que as consultas que antes não possuíam nenhum acerto, começam a ser capturadas somente pelas soluções por histograma média e modal.

O mesmo se verifica em relação a essas consultas no intervalo das cinqüenta páginas, figura 12. Nas demais consultas, em dez o *HITS* não apresenta nenhuma porcentagem de acerto e nas outras sete apresenta um desempenho inferior. Este varia de dez por cento até setenta por cento menos do que as outras duas soluções, levando em consideração os intervalos de dez, vinte e cinqüenta páginas.

Intervalo	Histograma Média	Histograma Modal	*HITS*
10	20	22,17391304	6,521739
20	34,7826087	37,39130435	9,565217
50	69,13043478	69,56521739	23,47826

Tabela 7: Média de acertos por intervalo de páginas para cada solução.

Na tabela 7 encontramos a média de acertos por solução em cada intervalo. Confirmando o que antes mencionado sobre o desempenho do *HITS*, as soluções por histograma média e modal apresentam um desempenho cerca de três a quatro vezes melhor nos intervalos. Entre as soluções por histograma média e modal, a solução modal possui um desempenho pouco melhor que a média, na média, não ficando atrás em nenhum intervalo.

A seguir apresentamos uma análise mais detalhada de dois dos oito tópicos antes definidos.

5.3. Tópico Fome

Nesse sub-tópico, passamos à análise da busca ao tópico fome comparativamente aos dois especialistas antes mencionados nesse trabalho.

5.3.1. Análise com a classificação do *Yahoo*

Primeiramente, na tabela 8, apresentamos as dez primeiras páginas que são retornadas pela máquina de busca do *Yahoo* (http://www.yahoo.com.br).

Posição	URL
1	http://www.clickfome.com.br/
2	http://www.fomezero.gov.br/
3	http://www.fomezero.org.br/
4	http://www.blogfome.blogger.com.br/
5	http://www.webciencia.com/13_fome.htm/
6	http://www.brasil.gov.br/
7	http://www.sitedafome.com.br/
8	http://www.acaodacidadania.com.br/
9	http://oglobo.globo.com/oglobo/especiais/fome/
10	http://www.fome.monashclubs.org/

Tabela 8: Classificação do *Yahoo*.

Com base nessa classificação, podemos observar, num gráfico compacto (figura 13) que reúne as dez páginas classificadas simultaneamente, a evolução da classificação dessas páginas, pelo algoritmo *XHITS*, ao longo da variação do parâmetro φ.

Podemos observar que a maioria das páginas começa a se aproximar rapidamente das primeiras colocações, para valores de φ inferiores a 0.1 e que, para valores de φ superiores a 0.1, se estabilizam.

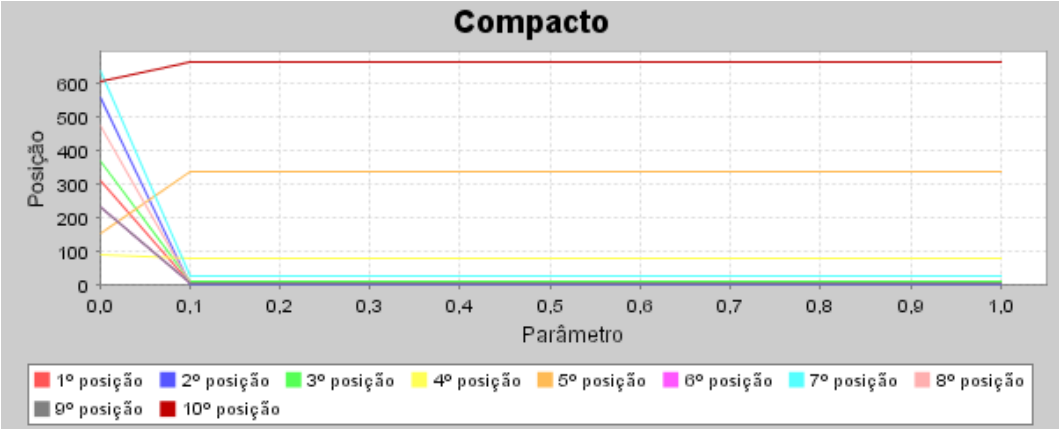

Figura 13: Gráfico variando o parâmetro φ de 0 a 1 com α = -1, β = 0, γ = 0 e θ = 0 pelo *XHITS*.

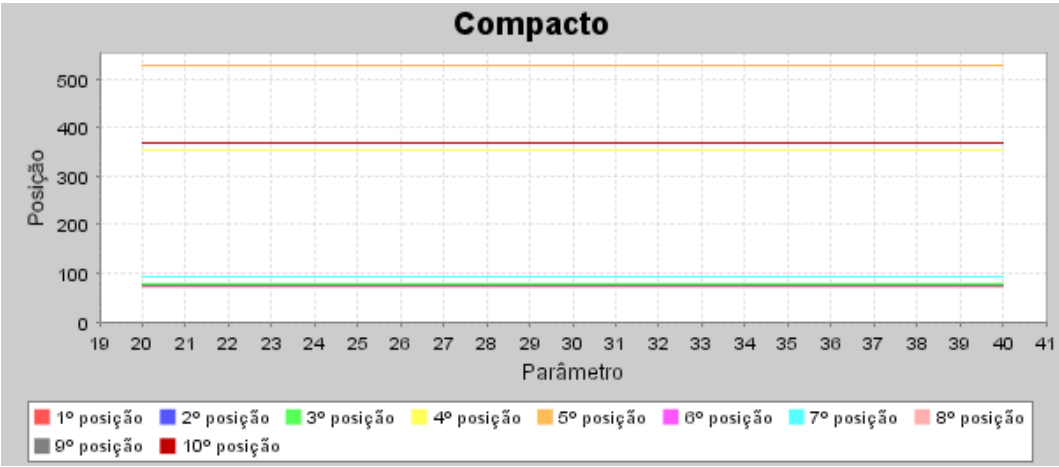

Figura 14: Gráfico variando o parâmetro de iteração do algoritmo entre 20 e 40 pelo *HITS*.

Na medida em que no algoritmo *HITS* não existe parâmetro a ser ajustado, o gráfico compacto da figura 14 representa a variação do número de iterações do algoritmo pela classificação das páginas. Verificamos que não há aproximação expressiva da classificação das páginas para as primeiras colocações, uma vez que a maior parte das páginas iniciam com classificações no intervalo de 50 a 550 e, ao longo da execução do algoritmo, não conseguem sair desse patamar.

Comparando os dois gráficos, verificamos que a aproximação do posicionamento das páginas conforme a classificação da máquina do *Yahoo* é claramente observada no gráfico do algoritmo *XHITS*. Para observar tal fato com maior precisão, os gráficos das figuras 15, 16 e 17, nos mostram os números de acertos das cinqüenta, vinte e dez primeiras páginas classificadas pelo *XHITS*, respectivamente.

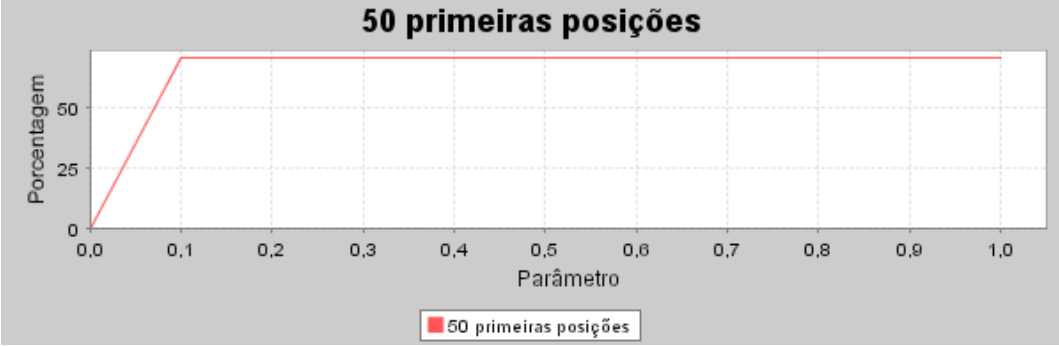

Figura 15: Número de acertos nas cinqüenta primeiras páginas pelo *XHITS*.

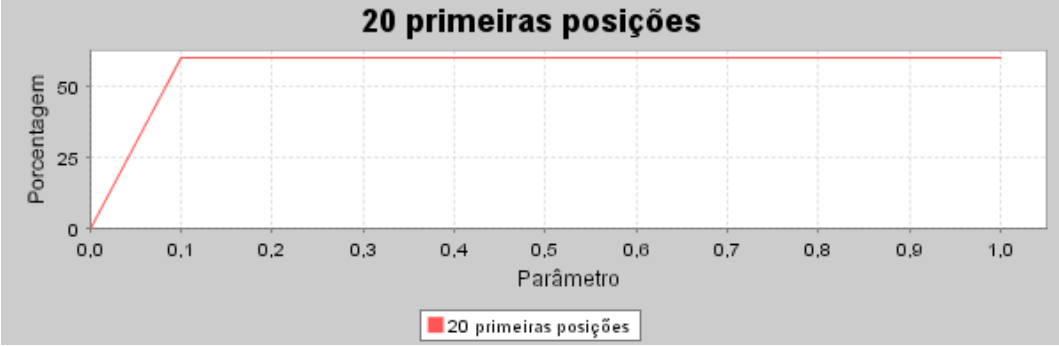

Figura 16: Número de acertos nas vinte primeiras páginas pelo *XHITS*.

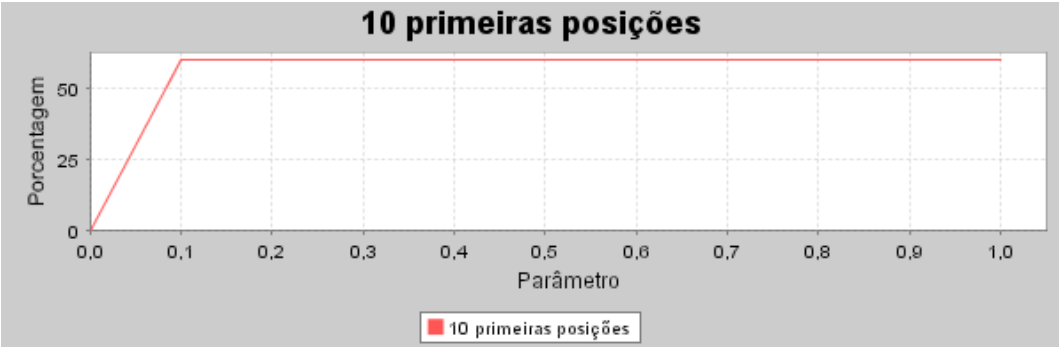

Figura 17: Número de acertos nas dez primeiras páginas pelo *XHITS*.

Nesse caso, o maior valor alcançado foi de sessenta por cento nas dez e vinte primeiras e setenta por cento nas cinqüenta primeiras páginas. Entretanto, com o algoritmo *HITS*, não conseguimos ajustar a classificação de nenhuma página, indicando um índice de acerto de zero por cento nas cinqüenta primeiras páginas (figura 18 e 19).

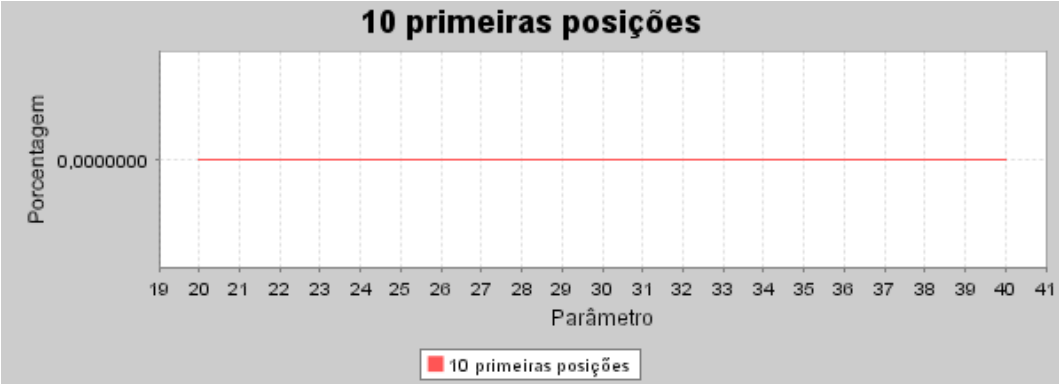

Figura 18: Número de acertos nas dez primeiras páginas pelo *HITS*.

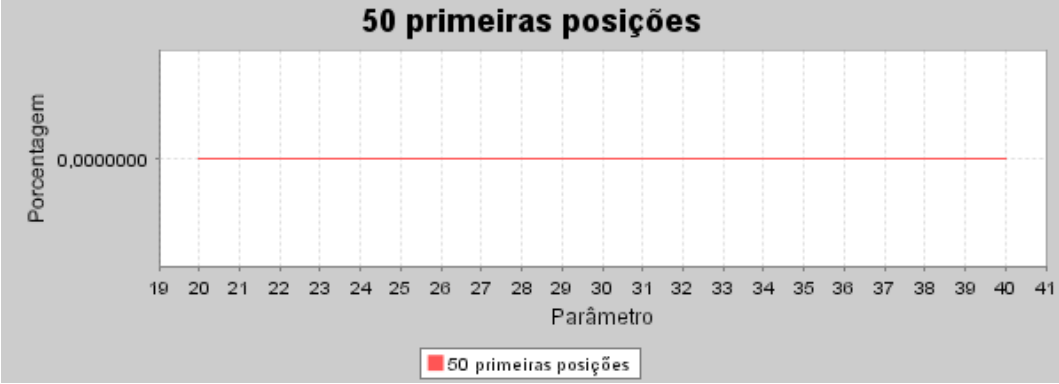

Figura 19: Número de acertos nas cinqüentas primeiras páginas pelo *HITS*.

Quanto à classificação das vinte primeiras páginas, podemos verificar, inicialmente, que o empate do valor das autoridades só acontece a partir da 11º página no *XHITS* (tabela 9) e a partir da 2º no *HITS* (tabela 10).

Rank	Url	Authority
1	http://www.fomezero.gov.br/	0.8022446352339896
2	http://luciana.misura.org/cat_me_deu_uma_fome.html/	0.5158554832683204
3	http://www.clickfome.com.br/	0.16360263685164414
4	http://www.acaodacidadania.com.br/	0.14087584326895003
5	http://www.brasil.gov.br/	0.057066482345703724
6	http://www.bancodealimentos.org.br/	0.050087741541824965
7	http://www.presidencia.gov.br/mesa/	0.04477725345222171
8	http://www.agricultura.gov.br/	0.03876535449471342

9	http://www.fomezero.org.br/	0.03780545741228512
10	http://www.unicef.org/brazil/	0.03435298644518449
11	http://www.desaparecidosbr.com/	0.03362103689831758
12	http://www.soudapaz.org/	0.03362103689831758
13	http://www.amigosdaescola.com.br/	0.03362103689831758
14	http://www.se.gov.br/	0.03362103689831758
15	http://www.unesco.org.br/	0.03362103689831758
16	http://www.aracaju.com/denuncia.php/	0.03362103689831758
17	http://www.clickdoe.com.br/	0.03362103689831758
18	http://www.ajudabrasil.com.br/	0.03362103689831758
19	http://www.filantropia.org/	0.03362103689831758
20	http://www.enkontraki.com.br/	0.03362103689831758

Tabela 9: Vinte primeiras páginas classificadas com α = -1, β = 0, γ = 0, φ = 1 e θ = 0 pelo *XHITS*.

Rank	Url	Authority
1	http://www.movabletype.org/	0.12215456452489971
2	http://www.mundopequeno.com/	0.12035856798702864
3	http://www.atl-turismolisboa.pt/	0.12035856798702864
4	http://www.evandromaciel.com/	0.12035856798702864
5	http://www.mcdonalds.com/countries/usa/whatsnew/salads/index.html/	0.12035856798702864
6	http://www.ilcaffediroma.pt/	0.12035856798702864
7	http://www.redlobster.com/	0.12035856798702864
8	http://www.danielsansao.com.br/motocontinuo/	0.12035856798702864
9	http://spacebabe.blogspot.com/	0.12035856798702864
10	http://www.carrabbas.com/	0.12035856798702864
11	http://www.obarquinho.com/wow/	0.12035856798702864
12	http://www.lu3.com/weblug/	0.12035856798702864
13	http://www.benihana.com/default.asp/	0.12035856798702864
14	http://www.espm.br/	0.12035856798702864
15	http://www.wendys.com/index.html/	0.12035856798702864
16	http://www.mel.blogger.com.br/	0.12035856798702864
17	http://www.google.com/search?hl=en&ie=UTF-8&oe=UTF-8&q=apple+varieties/	0.12035856798702864
18	http://global.mms.com/us/index.jsp/	0.12035856798702864
19	http://www.zoup.com/	0.12035856798702864
20	http://www.veja-rio.com.br/	0.12035856798702864

Tabela 10: Vinte primeiras páginas classificadas pelo *HITS*.

Esse trabalho não se propõe a avaliar a qualidade das páginas retornadas, porém podemos verificar que nas onze primeiras páginas retornadas pelo algoritmo *XHITS* (tabela 9), encontram-se páginas oficiais do governo e páginas mantidas por organizações não governamentais de notório conhecimento popular.

Por fim, a porcentagem de acerto que o algoritmo *XHITS* efetuou nas cinqüenta primeiras páginas foi de setenta por cento em relação a zero por cento ao mesmo intervalo classificado pelo *HITS*.

5.3.2.
Análise com a classificação do *Google*

Reiniciando, agora, com a classificação da máquina de busca do *Google* (http://www.google.com.br) (tabela 11), podemos observar, num gráfico compacto (figura 20) que reúne as dez páginas classificadas simultaneamente, a evolução da classificação dessas páginas pelo algoritmo *XHITS* ao longo da variação do parâmetro φ.

Posição	URL
1	http://www.fomezero.gov.br/
2	http://www.clickfome.com.br/
3	http://www.fomezero.org.br/
4	http://www.sitedafome.com.br/
5	http://www.aiquefome.com.br/
6	http://www.webciencia.com/13_fome.htm/
7	http://www.blogfome.blogger.com.br/
8	http://www.acaodacidadania.com.br/
9	http://confrontos.no.sapo.pt/page4.html/
10	http://www.josuedecastro.com.br/port/fome.html/

Tabela 11: Classificação do *Google*

Podemos observar, que a metade das páginas começa a se aproximar rapidamente das primeiras colocações, para valores de φ inferiores a 0.1 e que, para valores de φ superiores a 0.1, se estabilizam.

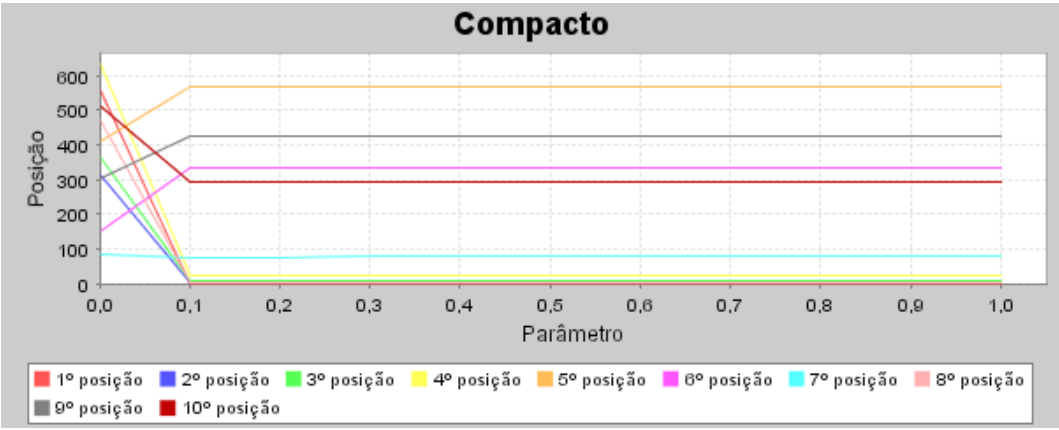

Figura 20: Gráfico variando o parâmetro φ de 0 a 1 com α = -1, β = 0, γ = 0 e θ = 0 pelo *XHITS*.

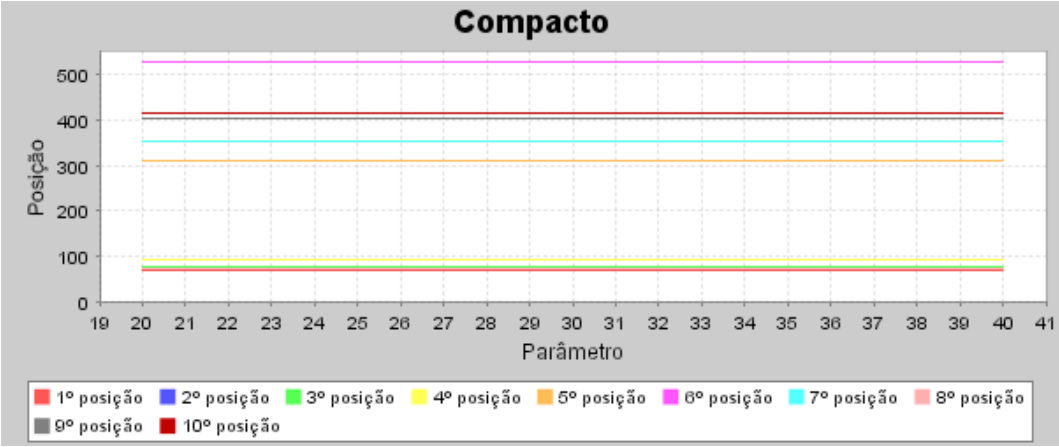

Figura 21: Gráfico variando o parâmetro de iteração do algoritmo entre 20 e 40 pelo *HITS*.

No algoritmo *HITS* não existe parâmetro a ser ajustado. O gráfico compacto da figura 21 representa a variação do número de iterações do algoritmo pela classificação das páginas. Podemos verificar que não há aproximação expressiva da classificação das páginas para as primeiras colocações, pois a maior parte das páginas inicia com classificações no intervalo de 50 a 600 e, ao longo da execução do algoritmo, não conseguem sair desse patamar.

Comparando os dois gráficos, verificamos que as páginas selecionadas para análise se aproximam mais intensamente das dez primeiras colocações no gráfico do algoritmo *XHITS*. Para observar tal fato com maior precisão, os gráficos das figuras 22, 23 e 24, mostram os números de acertos das cinqüenta, vinte e dez primeiras páginas classificadas pelo *XHITS*, respectivamente.

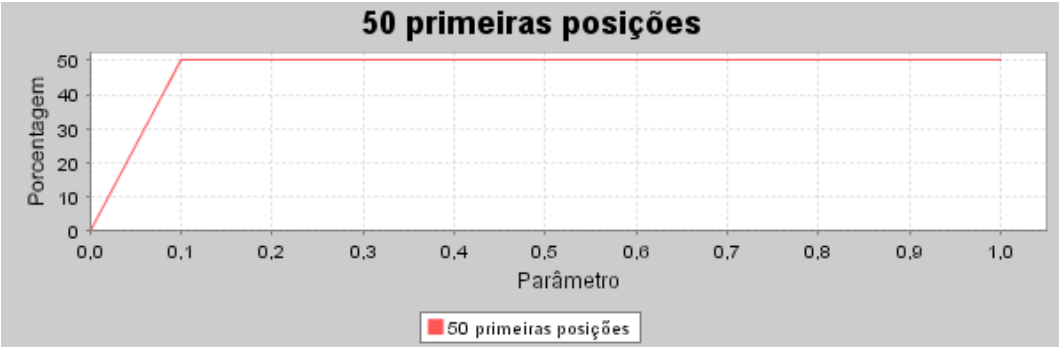

Figura 22: Número de acertos nas cinqüenta primeiras páginas pelo *XHITS*.

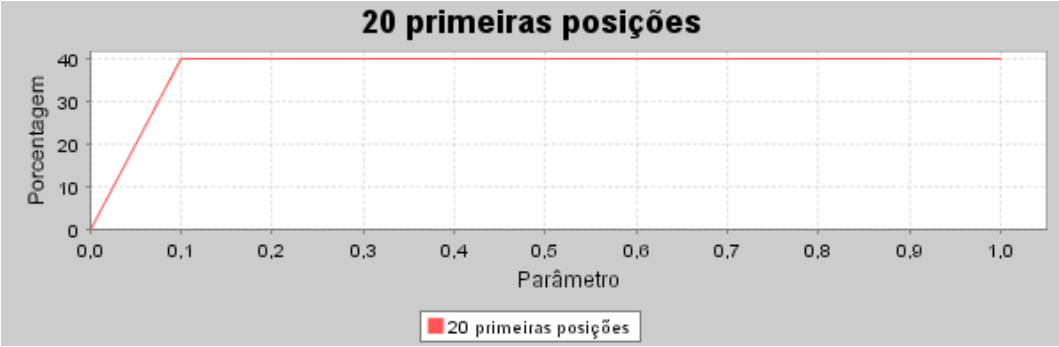

Figura 23: Número de acertos nas vinte primeiras páginas pelo *XHITS*.

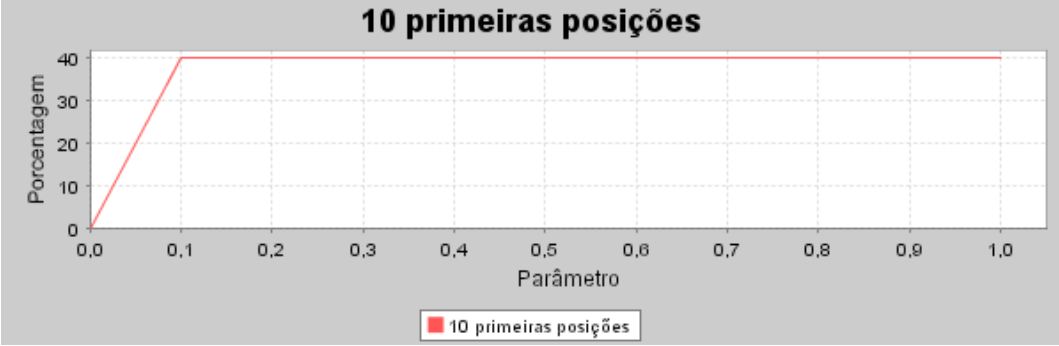

Figura 24: Número de acertos nas dez primeiras páginas pelo *XHITS*.

Nesse caso, o maior valor alcançado foi de quarenta por cento nas dez e vinte primeiras e cinqüenta por cento nas cinqüenta primeiras páginas. Entretanto, com o algoritmo *HITS*, não conseguimos ajustar a classificação de nenhuma página, indicando um índice de acerto de zero por cento nas cinqüenta primeiras páginas (figura 25 e 26).

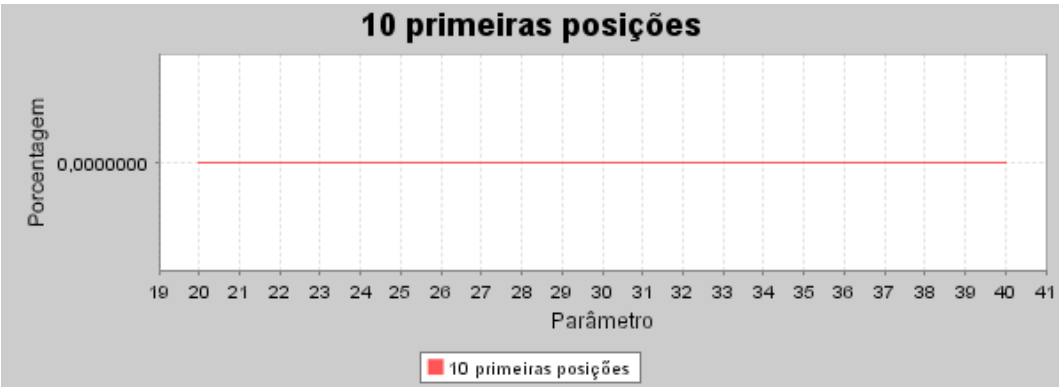

Figura 25: Número de acertos nas dez primeiras páginas pelo *HITS*.

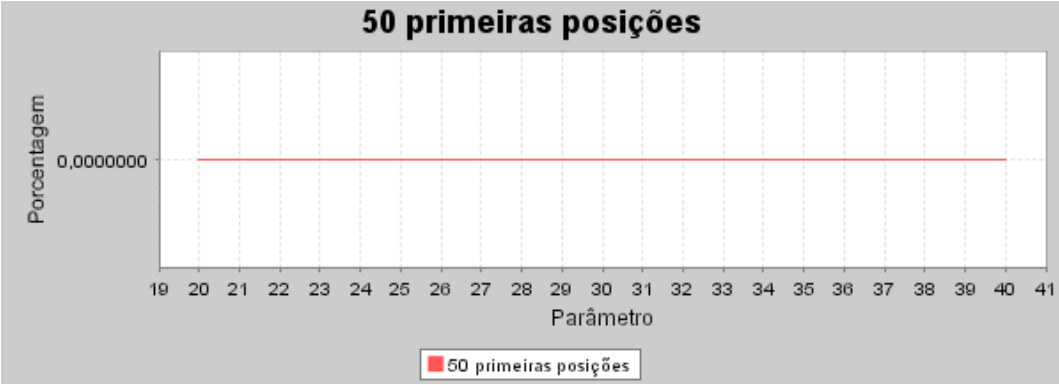

Figura 26: Número de acertos nas cinqüenta primeiras páginas pelo *HITS*.

Como os parâmetros são os mesmos utilizados no tópico anterior, α = -1, β = 0, γ = 0, φ = 1 e θ = 0, não há de se repetir os mesmos resultados.

Por fim, a porcentagem de acerto que o algoritmo *XHITS* efetuou nas cinqüenta primeiras páginas foi de cinqüenta por cento em relação a zero por cento ao mesmo intervalo classificado pelo *HITS*.

5.4. Tópico *Harvard*

Nesse sub-tópico analisamos a busca ao tópico *Harvard* comparativamente aos dois especialistas.

5.4.1. Análise com a classificação do *Yahoo*

Primeiramente, na tabela 12, vemos as dez primeiras páginas que são retornadas pela máquina de busca do *Yahoo*.

Posição	URL
1	http://www.harvard.edu/
2	http://www.law.harvard.edu/
3	http://lib.harvard.edu/
4	http://www.math.harvard.edu/
5	http://www.hms.harvard.edu/
6	http://blogs.law.harvard.edu/
7	http://www.harvard.edu/museums/
8	http://www.economics.harvard.edu/
9	http://www.hbs.edu/
10	http://cfa-www.harvard.edu/

Tabela 12: Classificação do *Yahoo*.

Podemos observar que a maioria das páginas começam a se aproximar rapidamente das primeiras colocações, para valores de θ inferiores a 0.7 e que, para valores de θ superiores a 0.7, se estabilizam.

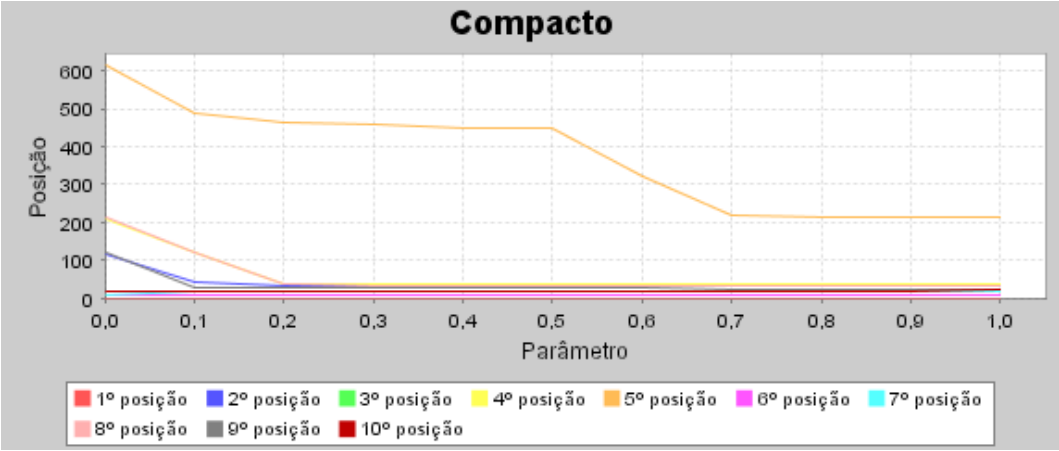

Figura 27: Gráfico variando o parâmetro θ de 0 a 1 com α = -1, β = -1, γ = 0 e φ = 0 pelo *XHITS*.

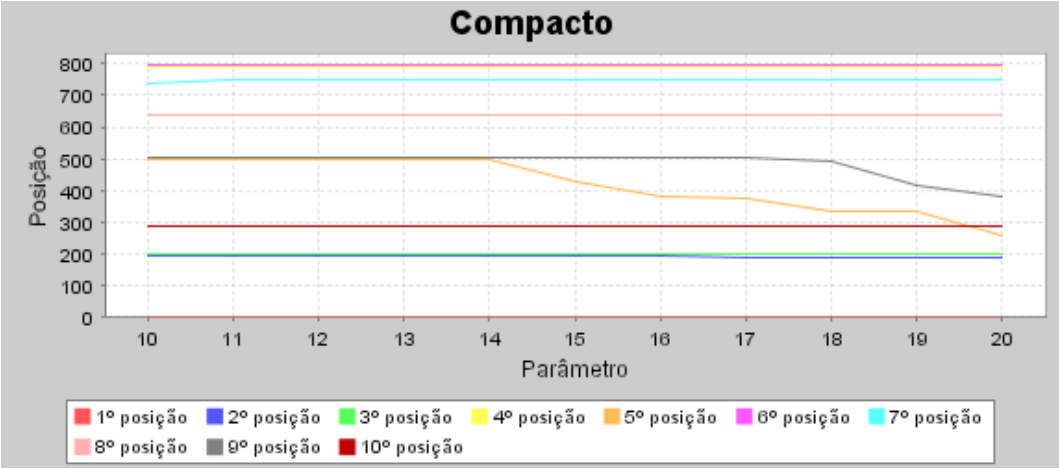

Figura 28: Gráfico variando o parâmetro de iteração do algoritmo entre 10 e 20 pelo *HITS*.

Já no algoritmo *HITS* podemos verificar que não há aproximação expressiva da classificação das páginas para as primeiras colocações, pois a maior parte das páginas iniciam com classificações no intervalo de 150 a 800 e, ao longo da execução do algoritmo, não conseguem sair desse patamar.

Comparando os dois gráficos, verificamos que as páginas selecionadas para análise se aproximam mais intensamente das dez primeiras colocações no gráfico do algoritmo *XHITS*. Para observar tal fato com maior precisão, os

gráficos das figuras 29, 30 e 31, nos mostram os números de acertos das cinqüenta, vinte e dez primeiras páginas classificadas pelo *XHITS*, respectivamente.

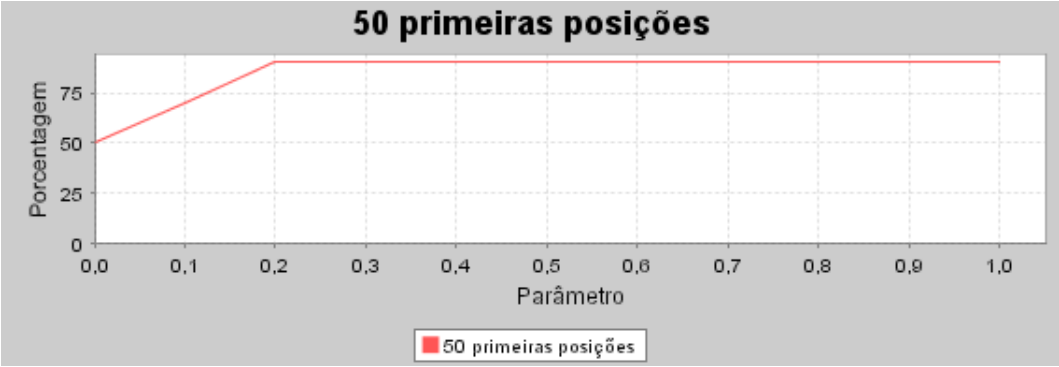

Figura 29: Número de acertos nas cinqüenta primeiras páginas pelo *XHITS*.

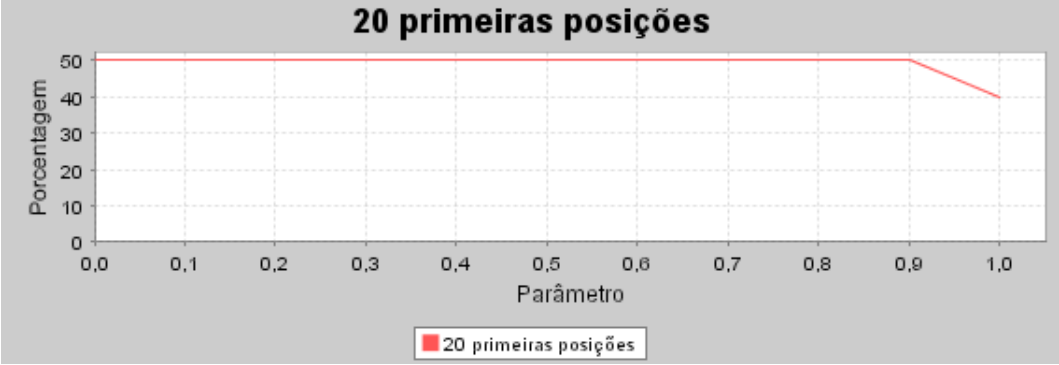

Figura 30: Número de acertos nas vinte primeiras páginas pelo *XHITS*.

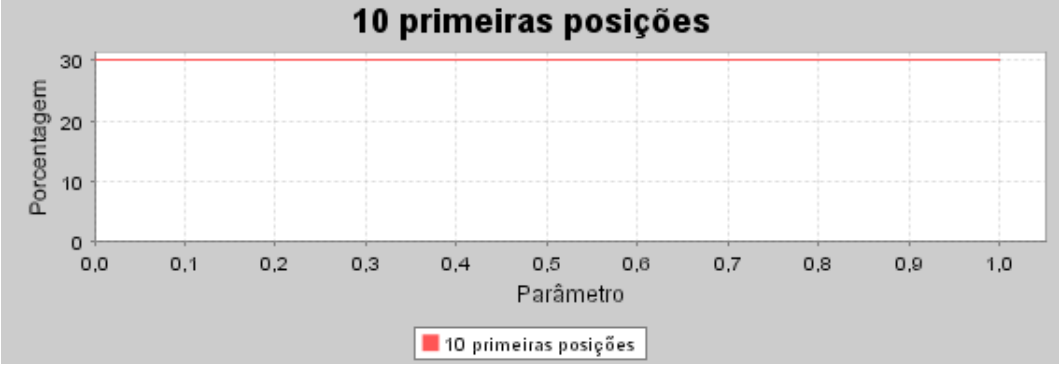

Figura 31: Número de acertos nas dez primeiras páginas pelo *XHITS*.

Nesse caso, o maior valor alcançado foi de trinta por cento nas dez primeiras, cinqüenta por cento por cento nas vinte primeiras e oitenta por cento nas cinqüenta primeiras páginas. Entretanto, com o algoritmo *HITS*, apenas

conseguimos ajustar a classificação da primeira página, indicando um índice de acerto apenas de dez por cento nas cem primeiras páginas (figura 32 e 33).

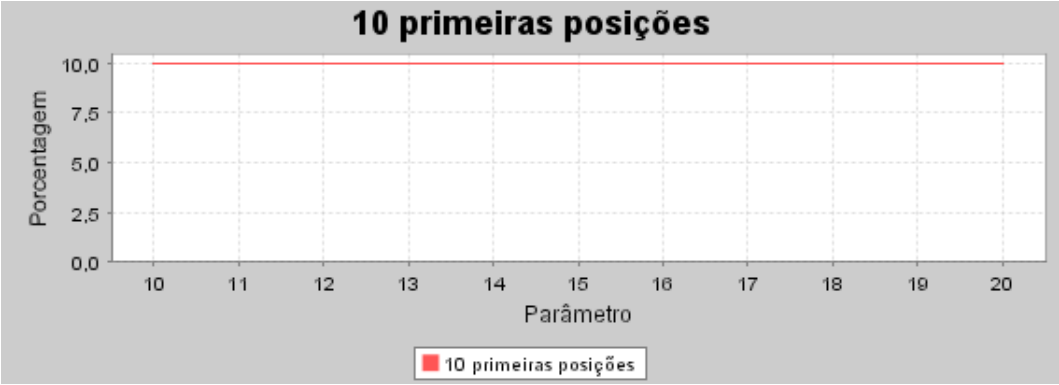

Figura 32: Número de acertos nas dez primeiras páginas pelo *HITS*.

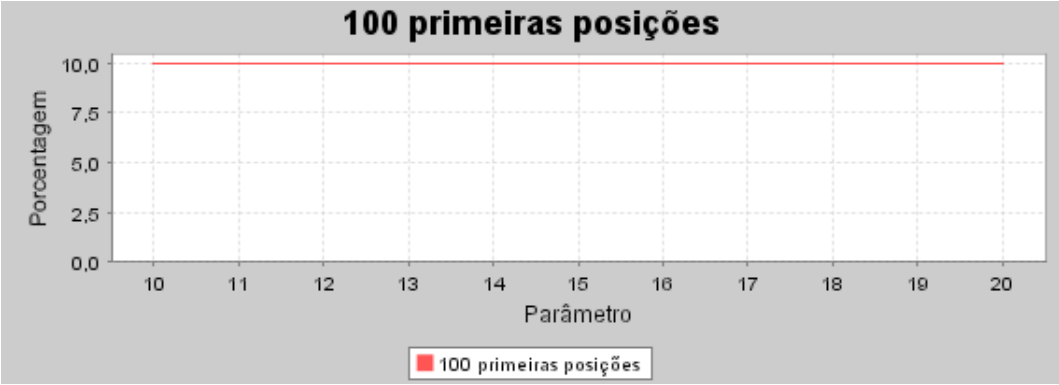

Figura 33: Número de acertos nas cem primeiras páginas pelo *HITS*.

Quanto à classificação das vinte primeiras páginas, podemos verificar que o empate do valor das autoridades não acontece no *XHITS* (tabela 13) e ocorre a partir da 2º página no *HITS* (tabela 14).

Rank	Url	Authority
1	http://www.harvard.edu/	0.7328648267850641
2	http://www.worldhealthnews.harvard.edu/	0.5193583640475491
3	http://gseweb.harvard.edu/	0.12169438409504339
4	http://www.mcb.harvard.edu/BioLinks.html/	0.11386946929472244
5	http://www.gse.harvard.edu/	0.08743908099177912
6	http://www.hsph.harvard.edu/	0.07471802811971563
7	http://www.edletter.org/	0.06792062939150717
8	http://gseweb.harvard.edu/~hepg/her.html/	0.05743517285193597
9	http://lib.harvard.edu/	0.05463429037482291
10	http://adswww.harvard.edu/	0.04999647832803369
11	http://blogs.law.harvard.edu/	0.04981874229554323
12	http://www.fas.harvard.edu/home/	0.048935028683474706
13	http://www.deas.harvard.edu/	0.04830900346644848
14	http://www.emergency.harvard.edu/	0.04531105300751359
15	http://www.cid.harvard.edu/	0.04506753010077985

16	http://www.peabody.harvard.edu/	0.04097443991711938
17	http://www.harvard.edu/museums/	0.039340409219443394
18	http://holliscatalog.harvard.edu/	0.03889362650808303
19	http://www.ksg.harvard.edu/	0.03657112726987767
20	http://www.hiid.harvard.edu/	0.03640305793543966

Tabela 13: Vinte primeiras páginas classificadas com α = 0.3, β = 0, γ = 0, φ = 1 e θ = 0.6 pelo *XHITS*.

Rank	Url	Authority
1	http://www.harvard.edu/	0.1252340930078445
2	http://www.hsph.harvard.edu/	0.11121026697413418
3	http://www.latimes.com/news/printedition/front/la-sci-stemcells24jan24,1,5876765.story?coll=la-headlines-frontpage/	0.09631338514338517
4	http://www.guardian.co.uk/food/Story/0,2763,1395335,00.html/	0.09631338514338517
5	http://www.nytimes.com/2005/01/18/health/nutrition/18cons.html/	0.09631338514338517
6	http://www.chicagotribune.com/features/lifestyle/health/chi-0501160418jan16,1,3271040.story?coll=chi-homepagenews2-utl/	0.09631338514338517
7	http://www.japantimes.com/cgi-bin/getarticle.pl5?nn20050119a2.htm/	0.09631338514338517
8	http://www.nytimes.com/2005/01/18/opinion/18tues2.html/	0.09631338514338517
9	http://www.nytimes.com/2005/01/20/politics/20cabinet.html/	0.09631338514338517
10	http://www.boston.com/news/local/massachusetts/articles/2005/01/21/infections_not_listed_in_bu_bid_for_biolab/	0.09631338514338517
11	http://news.yahoo.com/news?tmpl=story&u=/ap/20050124/ap_on_he_me/bird_flu_1/	0.09631338514338517
12	http://abcnews.go.com/Health/wireStory?id=437444/	0.09631338514338517
13	http://www.chicagotribune.com/features/lifestyle/health/chi-0501190365jan19,1,6023559.story?coll=chi-business-hed/	0.09631338514338517
14	http://sfgate.com/cgi-bin/article.cgi?file=/chronicle/archive/2005/01/18/EDGT7AQRRK1.DTL/	0.09631338514338517
15	http://www.csmonitor.com/2005/0119/p14s02-lifo.html/	0.09631338514338517
16	http://www.sacbee.com/content/news/story/12069692p-12939907c.html/	0.09631338514338517
17	http://www.latimes.com/news/printedition/asection/la-na-planes20jan20,1,6565419.story?coll=la-news-a_section/	0.09631338514338517
18	http://www.sacbee.com/content/politics/ca/story/12095894p-12966018c.html/	0.09631338514338517
19	http://www.boston.com/news/local/massachusetts/articles/2005/01/21/probe_of_bu_lab_illnesses_looks_to_a_lurking_contaminant/	0.09631338514338517
20	http://www.smh.com.au/news/National/Human-trials-for-cancer-HIV-cure/2005/01/15/1105582768080.html/	0.09631338514338517

Tabela 14: Vinte primeiras páginas classificadas pelo *HITS*.

De acordo com *Jon Kleinberg* [1,2], as maiores autoridades sobre o tópico em tela seriam as páginas relacionadas com a Universidade de *Harvard*.

Analisando a tabela 13, podemos verificar que das vinte primeiras páginas, dezenove são autoridades nas diversas áreas de atuação de *Harvard*: departamento de Direito, departamento de Biologia Molecular e Celular etc.

Ao contrário das páginas retornadas pelo algoritmo *XHITS*, as páginas retornadas pelo algoritmo *HITS* (tabela 14), com exceção das duas primeiras, são de notícias e citações sobre *Harvard*.

Por fim, a porcentagem de acerto que o algoritmo *XHITS* efetuou nas cinqüenta primeiras páginas foi de setenta por cento em relação a zero por cento ao mesmo intervalo classificado pelo *HITS*.

5.4.2. Análise com a classificação do *Google*

Inicialmente, na tabela 15, vemos as dez primeiras páginas que são retornadas pela máquina de busca do *Google*.

Posição	URL
1	http://www.harvard.edu/
2	http://www.hbs.edu/
3	http://www.law.harvard.edu/
4	http://www.thecrimson.com/
5	http://www.harvardlawreview.org/
6	http://harvardbusinessonline.hbsp.harvard.edu/b02/en/hbr/hbr_home.jhtml/
7	http://www.hms.harvard.edu/
8	http://www.intelihealth.com/
9	http://www.joslin.org/
10	http://www.hbsp.harvard.edu/

Tabela 15: Classificação do *Google*.

Podemos observar, que neste caso a maioria das páginas não se aproxima das primeiras colocações. A posição das páginas se estabilizam para valores de θ superiores a 0.7.

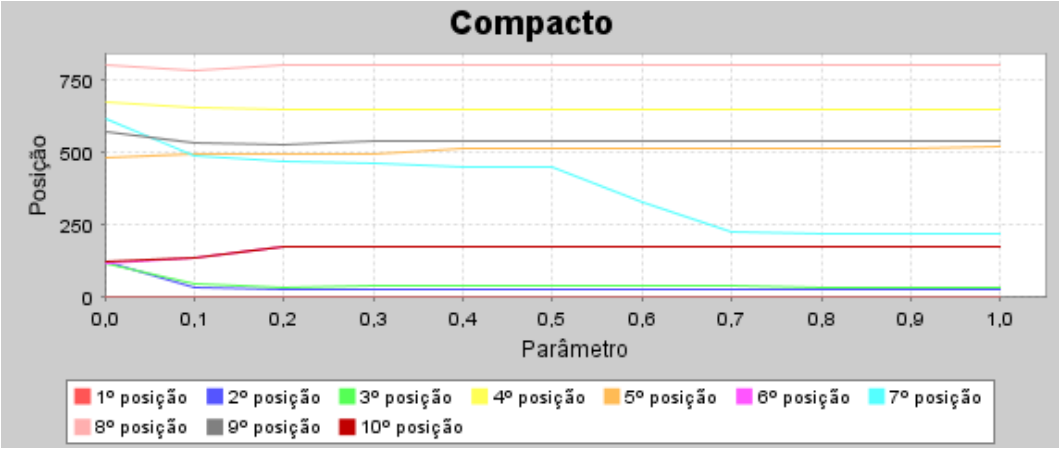

Figura 34: Gráfico variando o parâmetro θ de 0 a 1 com α = 0.3, β = 0, γ = 0 e φ = 1 pelo *XHITS*.

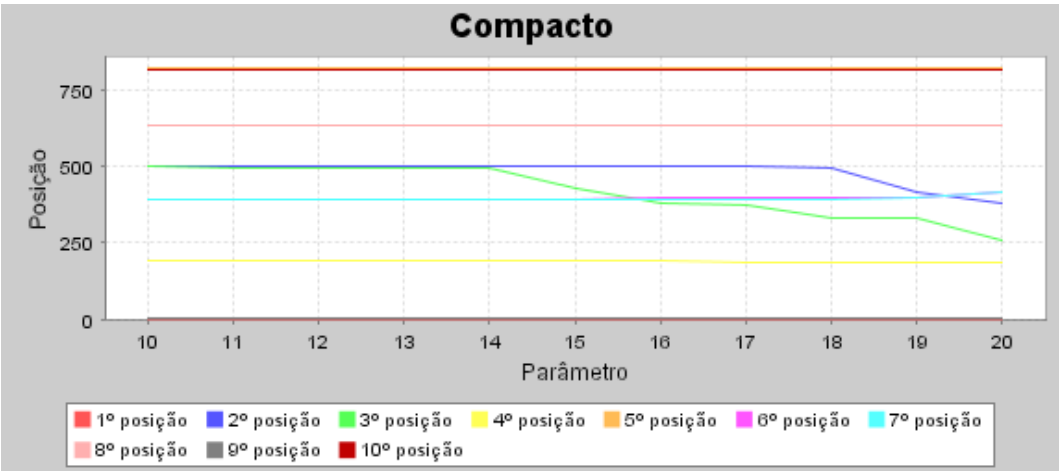

Figura 35: Gráfico variando o parâmetro de iteração do algoritmo entre 10 e 20 pelo *HITS*.

Neste caso, com a classificação da máquina de busca do *Google*, o *XHITS* não consegue atuar de forma a aproximar as classificações. O mesmo acontece com o algoritmo *HITS.* Não há aproximação expressiva da classificação das páginas para as primeiras colocações (figuras 34 e 35).

Comparando os dois gráficos, verificamos que nenhum dos dois aproxima a maior parte das páginas em análise das dez primeiras posições. Para observar tal fato com maior precisão, os gráficos das figuras 36, 37 e 38, nos mostram os números de acertos das cinqüenta, vinte e dez primeiras páginas classificadas pelo *XHITS*, respectivamente.

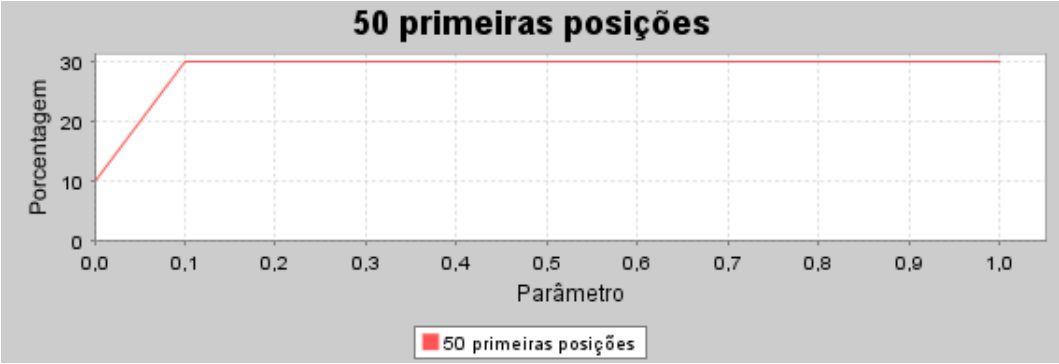

Figura 36: Número de acertos nas cinqüenta primeiras páginas pelo *XHITS*.

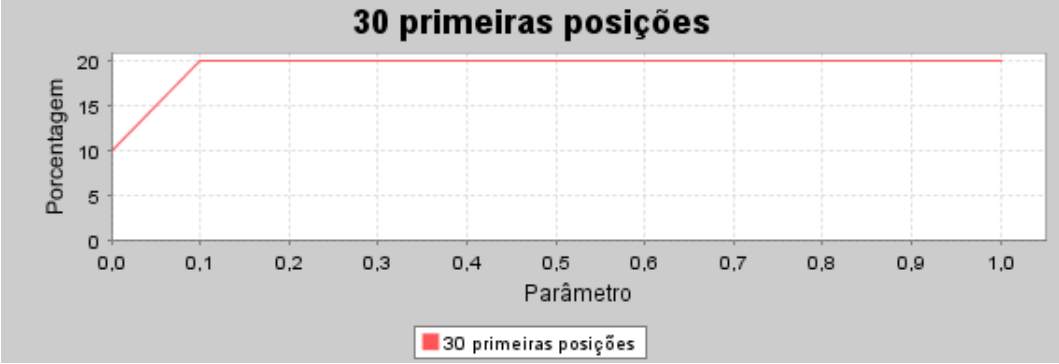

Figura 37: Número de acertos nas trinta primeiras páginas pelo *XHITS*.

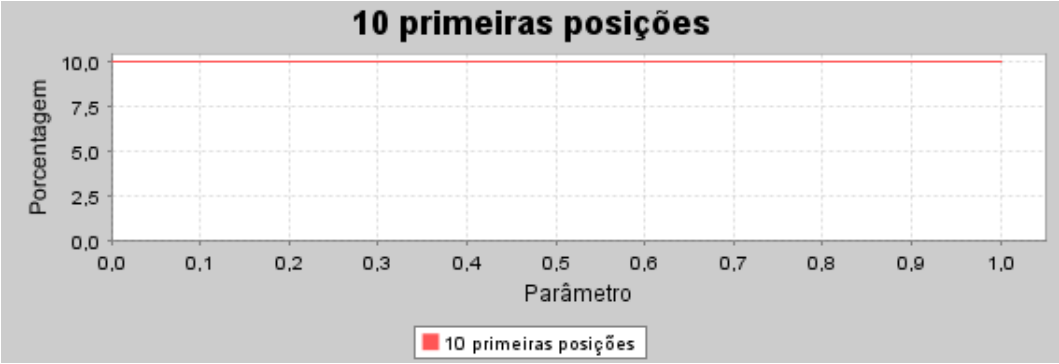

Figura 38: Número de acertos nas dez primeiras páginas pelo *XHITS*.

Nesse caso, o maior valor alcançado foi de dez por cento nas dez primeiras, vinte por cento nas trinta primeiras e trinta por cento nas cinqüenta primeiras páginas. Entretanto, com o algoritmo *HITS*, apenas conseguimos ajustar a classificação de duas páginas, indicando um índice de acerto de vinte por cento nas cem primeiras páginas (figura 39 e 40).

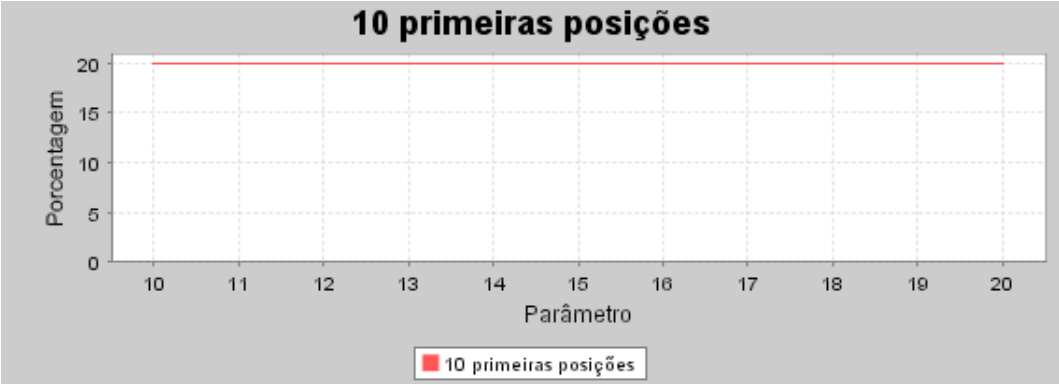

Figura 39: Número de acertos nas dez primeiras páginas pelo *HITS*.

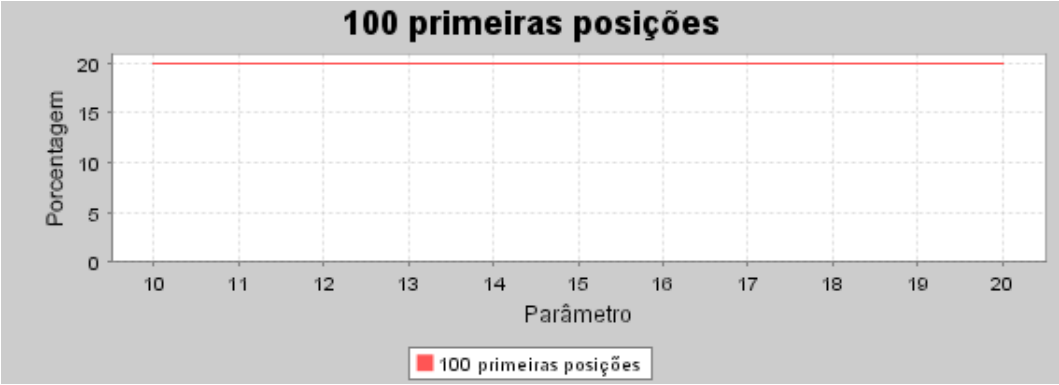

Figura 40: Número de acertos nas cem primeiras páginas pelo *HITS*.

O baixo desempenho aqui demonstrado pode ser justificado pelas páginas que foram retornadas pela máquina de busca do *Google*. Um exemplo é a *url* http://www.joslin.org. Esta página é um centro de informações sobre diabetes e possui apenas uma afiliação com a escola de medicina de *Harvard*. Neste caso, analisando a classificação do *Yahoo*, vemos que as páginas retornadas, em sua maioria, estão vinculadas a algum setor da Universidade. Não é estranho algumas páginas, como http://www.joslin.org, serem retornadas pelas máquinas de busca comerciais. Tais máquinas, por força do mercado, possuem mecanismos que alteram as classificações de forma a beneficiar uma ou outra página.

Por fim, a porcentagem de acerto que o algoritmo *XHITS* efetuou nas cinqüenta primeiras páginas foi de trinta por cento em relação a vinte por cento ao mesmo intervalo classificado pelo *HITS*.