RI – Relatório do Trabalho Prático 3

Danilo Ferreira e Silva[1]

[1]Departamento de Ciência da Computação – UFMG

danilofs@dcc.ufmg.br

1. Introdução

O objetivo deste trabalho foi a melhoria do processador de consulta do trabalho 2, introduzindo no raking dos documentos duas novas fontes de evidência, o *Pagerank* e texto das âncoras. As melhorias foram introduzidas e avaliadas com o mesmo conjunto de consultas disponíveis, onde foram constatadas melhorias em consultas navegacionais, mas também alguma piora em outros tipos de consulta.

O sistema está disponível para avaliação em:

`http://icse.labsoft.dcc.ufmg.br:9000/`

A URL só é acessível quando autenticado na rede interna do DCC (É possível usar a VPN do DCC para acessar de qualquer lugar).

2. Decisões de implementação

O indexador desenvolvido no TP1 foi modificado para extrair as âncoras dos documentos da coleção. Dentre estas âncoras, foram consideradas apenas aquelas que apontavam para uma *URL* contida dentro da própria coleção. Estas âncoras formam então ligações em um grafo da rede, em cima do qual foi calculado o *Pagerank*.

A tabela 1 mostra algumas estatísticas das âncoras coletadas. É interessante notar aproximadamente 15% dos documentos possuem algum link para os mesmos. Além disso, embora a maioria dos documentos tenha poucas referências, existem alguns com um número muito grande de referências, chegando até um máximo de 41184.

Tabela 1. Estatísticas da extração de âncoras

Métrica	Valor
Total de âncoras	34069647
Âncoras entre documentos da coleção	956234
Grau máximo de entrada de um documento	41184
Grau máximo de saída de um documento	59
Total de documentos	945642
Documentos referenciados	140548

2.1. Extração dos textos âncora

O texto contido nas *tags* âncora foram utilizados para gerar um novo arquivo de listas invertidas, associando os termos do mesmo ao documento referenciado. Além dos textos âncora, foram inseridos neste índice os termos do título de cada página. O motivo de tal decisão foi que muitas páginas da coleção não são referenciadas por qualquer outra. O objetivo eta que toda página tivesse algumas palavras chaves que a descrevessem, mesmo na ausência de âncoras.

2.2. Cálculo do *Pagerank*

Além dos textos das âncoras, elas foram usadas para contruir um grafo de ligações entre páginas. Neste grafo, cada vértice é uma página e ele armazena uma lista de páginas que apontam para ele.

Com estas informações, foi possível calcular o *Pagerank* de forma iterativa, baseando-se na seguinte equação:

$$PR(u,i) = q + (1-q) \sum_{v \in B_u} \frac{PR(v, i-1)}{L(v)}$$

Onde q é uma constante, i é a iteração, u e v são documentos, B_u são os documentos que apontam para u e $L(v)$ é a quantidade de documentos que v referencia. Para inicialização, foi usado $PR(u, 0) = 1$.

Os valores do *Pagerank* foram calculados em 50 iterações, onde foi constatado que os valores entre duas iterações já variavam muito pouco. Na figura 1 temos um gráfico do log do erro (*mean squared error*) por iteração, onde podemos observar a convergência do *Pagerank*. Quanto maior o valor de q, mais rápido é a convergência, como esperado, já que a contribuição do *Pagerank* dos vizinhos tem peso menor.

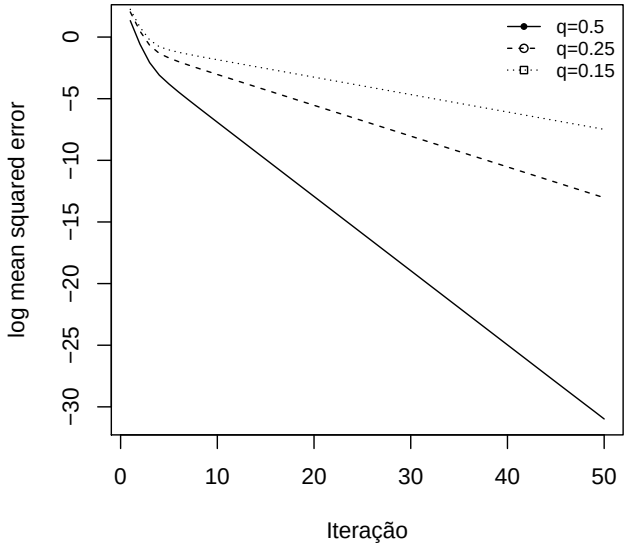


Figura 1. Convergência do *Pagerank*

2.3. Modelos de *ranking*

Com base nessas novas fontes de evidência, foram implementados novos modelos de ranking baseados no *BM25*:

bm25-anchorOnly Modelo BM25 levando em conta apenas ocorrências em texto âncora e título dos documentos.

bm25-anchor Modelo BM25 levando em conta todo o documento, incluindo texto âncora.

bm25-anchor-pr Modelo BM25 com texto âncora e fator multiplicador do documento baseado no log do *Pagerank*.

No modelo com a introdução do *Pagerank*, foi observado que páginas com um número baixo de referências causavam uma perturbação da ordenação que prejudicava o resultado. Para solucionar este problema foram desconsiderados os valores de *Pagerank* de páginas com um número de referências menor que um limiar.

3. Avaliação

Foram executadas todas as consultas de referência nos três modelos implementados e também no modelo BM25 puro para fins de comparação. As curvas de precisão por revocação são apresentadas nas figuras a seguir.

Uma primeira observação interessante é que o modelo *bm25-anchorOnly*, que ignora os termos do corpo do documento, em muitos casos foi superior ao *bm25*. Este resultado é de certa forma surpreendente, mas pode ser explicado pela natureza das consultas de teste, pois várias delas contém poucos termos, são muito gerais ou de caráter navegacional.

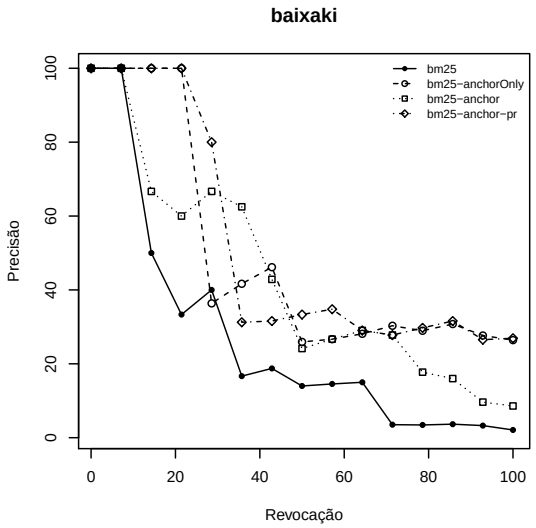
O modelo *bm25-anchor*, que combina o conteúdo do documento e texto âncora, foi levemente superior na maioria dos casos, o que era esperado.

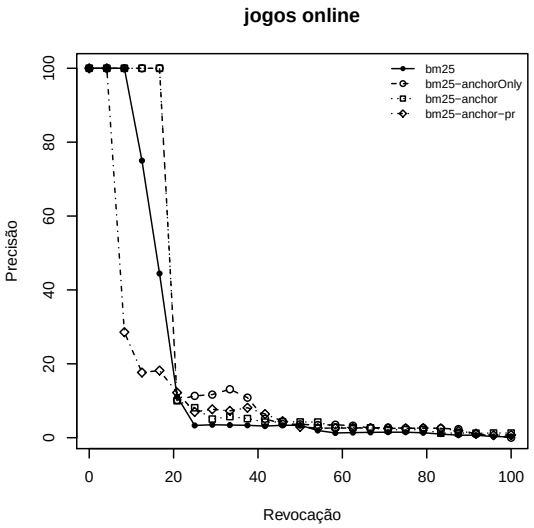
Os resultados obtidos com a introdução do *Pagerank*, surpreendentemente, não foram bons. Consultas como *globo* ou *terra* tiveram resultados bem superiores com a introdução do *Pagerank*. No entanto, consultas como *receita federal* e *record*, que também são navegacionais, tiveram seus resultados prejudicados pela introdução do mesmo.

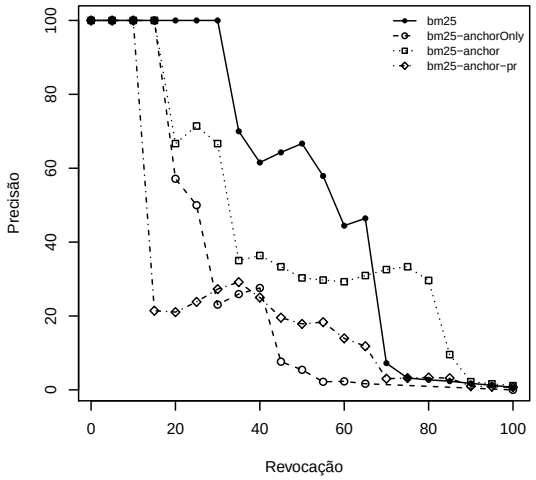
Duas hipóteses foram levantadas que podem explicar estes resultados, mas precisariam ser melhor estudadas para que fossem confirmadas. A primeira delas é que a coleção utilizada é uma amostragem pequena da Web, de forma que a esparsidade das ligações entre documentos da coleção torna o *Pagerank* pouco realístico.

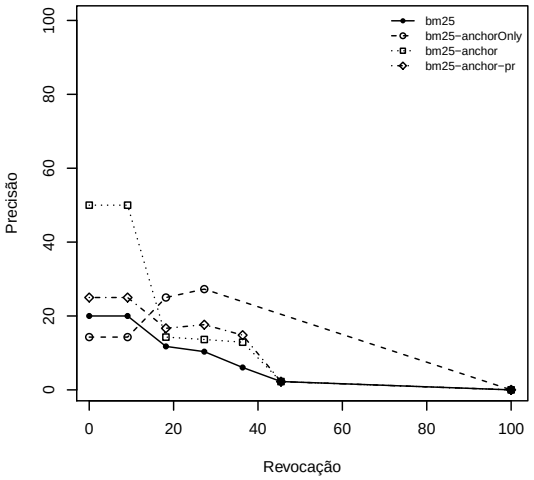
A segunda diz respeito ao fato que âncoras cujo endereço não fosse uma *URL* completa foram desconsideradas. Ou seja, *links* relativos, muito comuns para navegação interna dentro de um *site*, não foram contabilizados. Estes links podem ter um papel importante na transferência de *Pagerank* para a página principal.

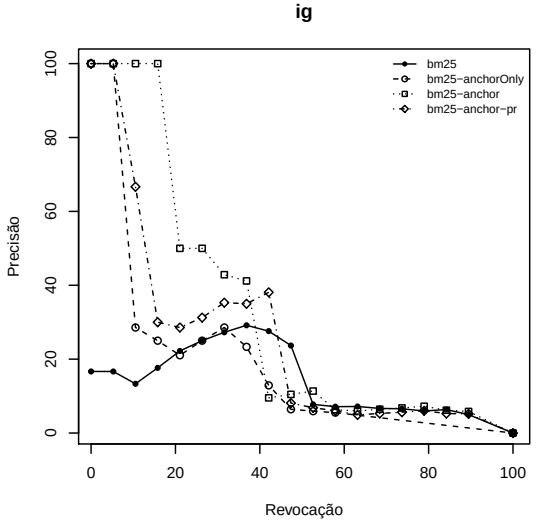
Além disso, uma última observação surpreendente foi o fato que as consultas como *mario*, *frases de amor* e pânico tiveram resultados amplamentes superiores no model BM25 puro.

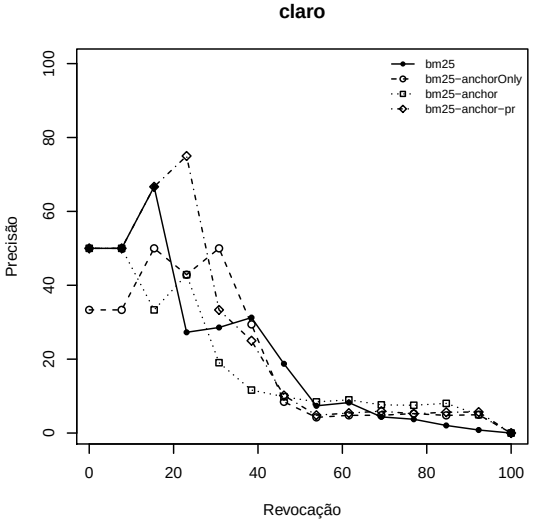

baixaki
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

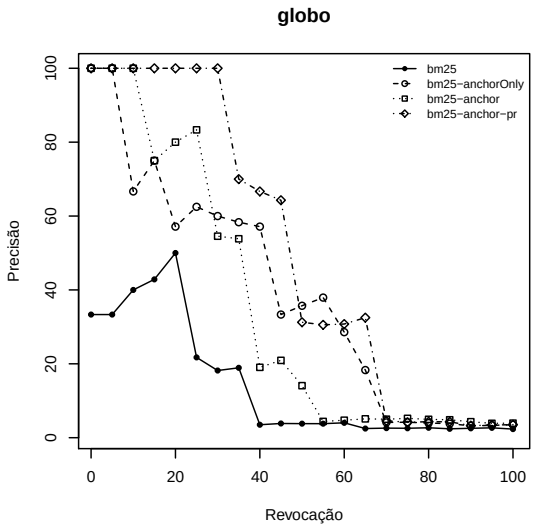

jogos online
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

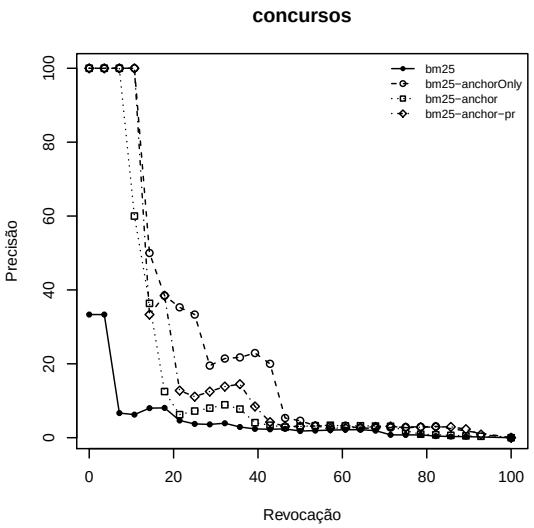

receita federal
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

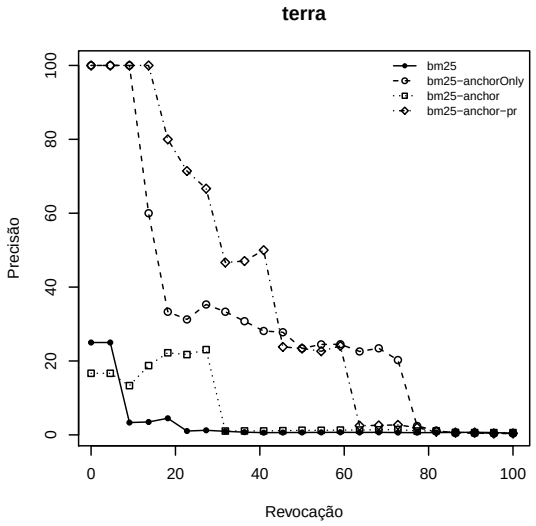

poquer
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

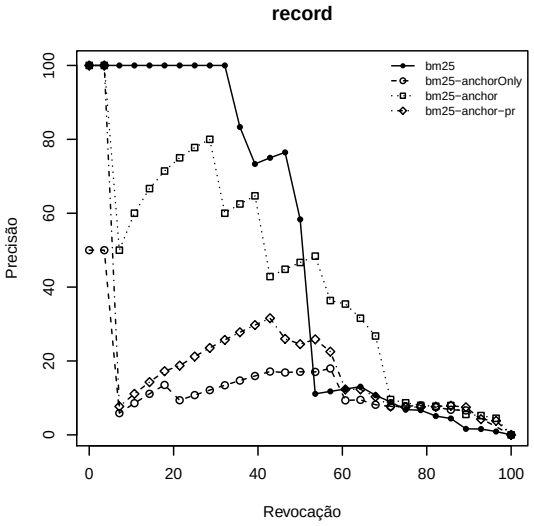

ig
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

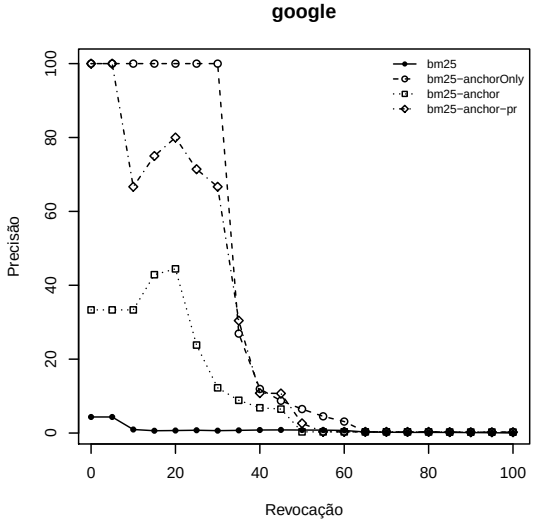

claro
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

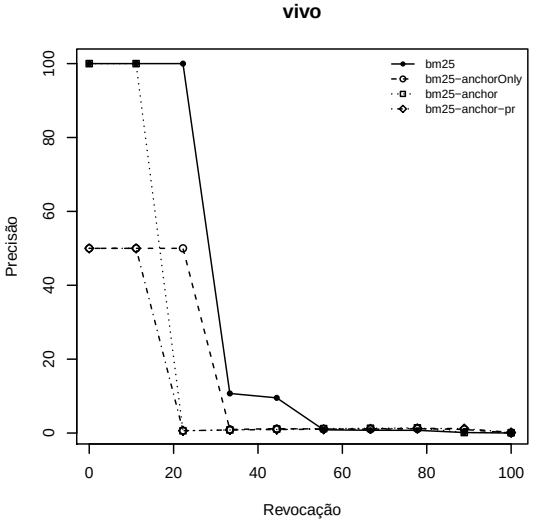

globo
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

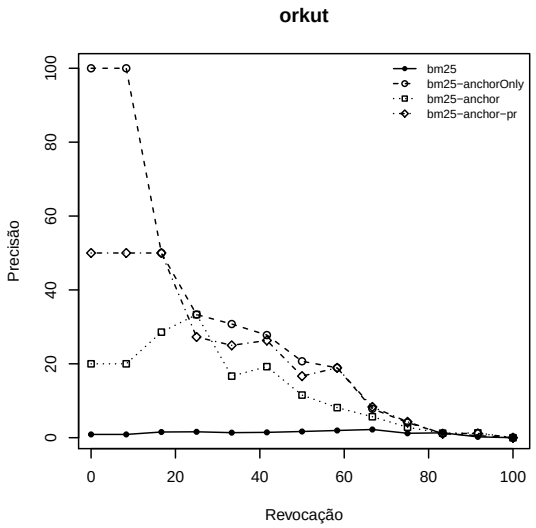

concursos
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

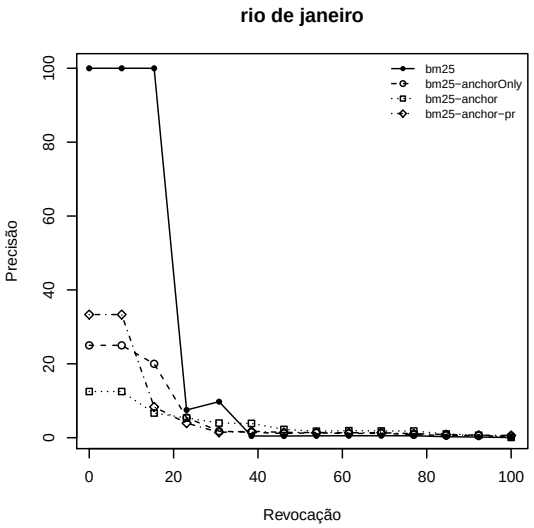

terra
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

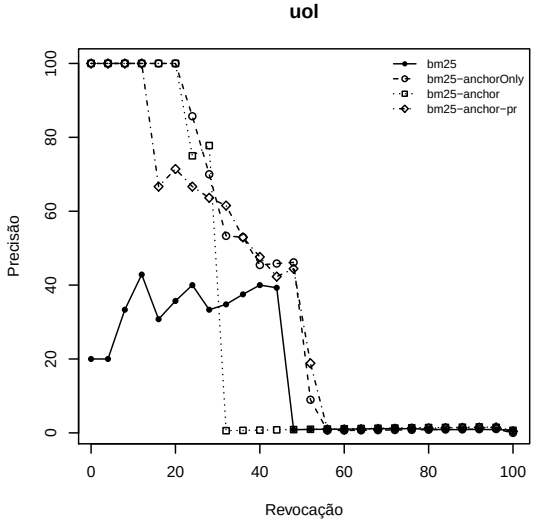

record
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

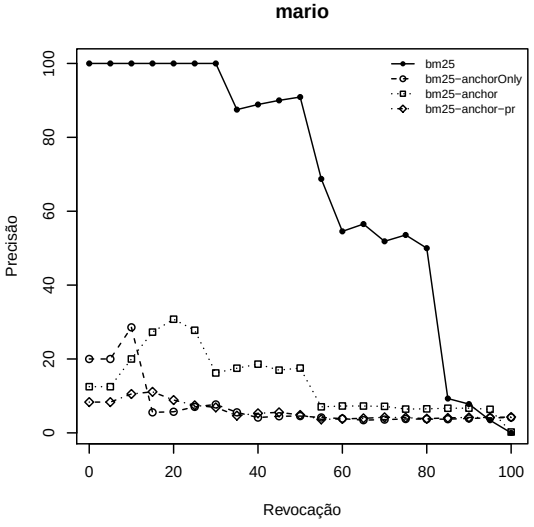

google
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

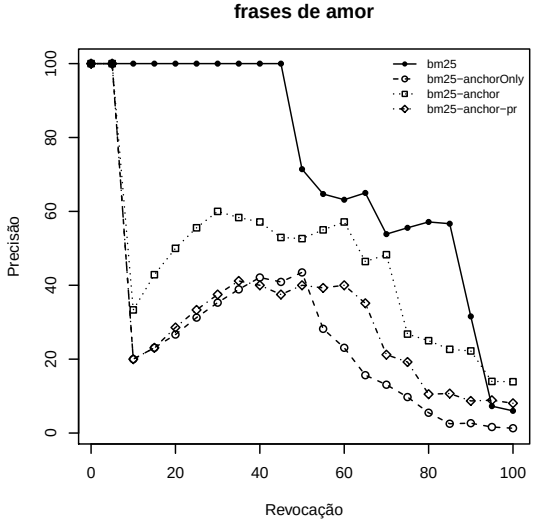

vivo
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

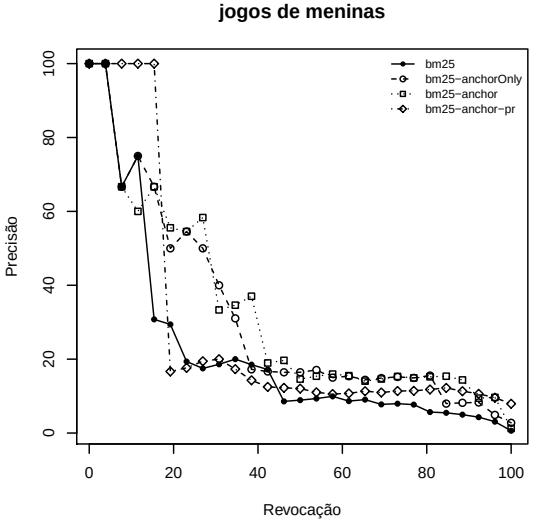

orkut
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

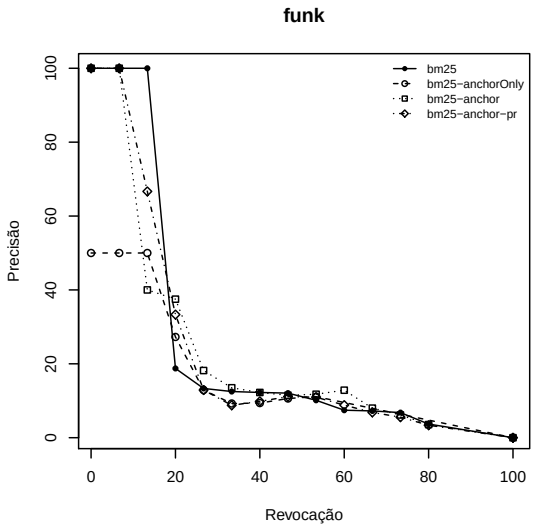

rio de janeiro
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

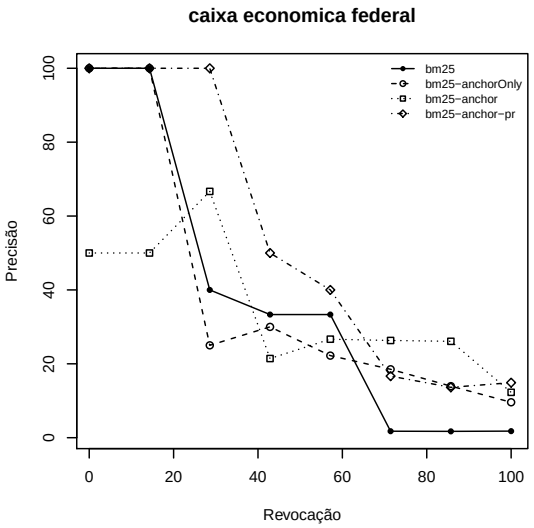

uol
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

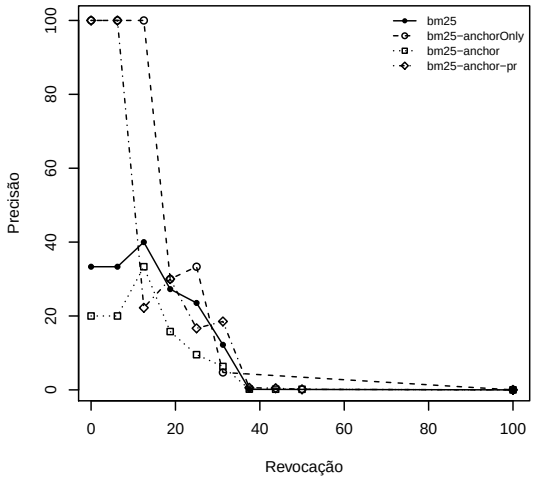

mario
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

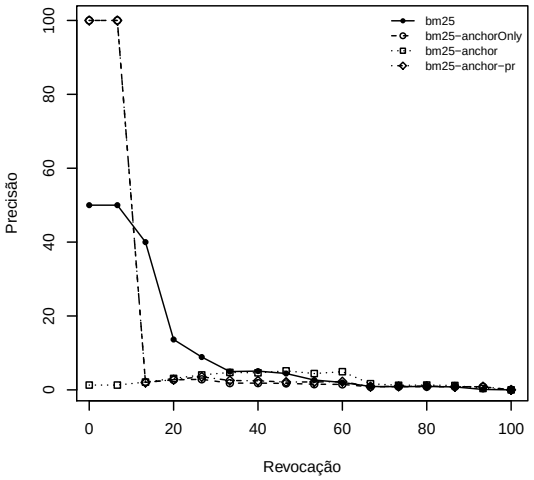

frases de amor
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

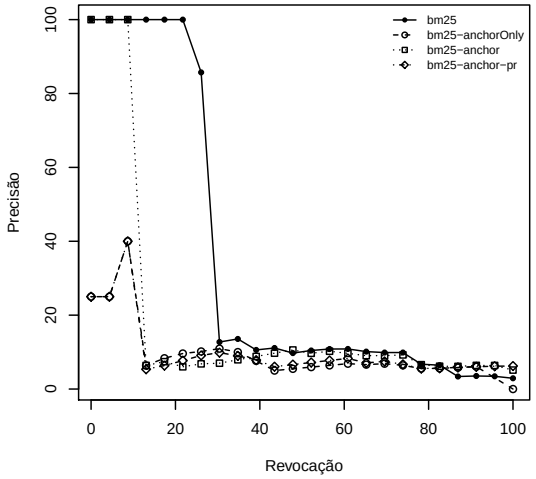

jogos de meninas
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

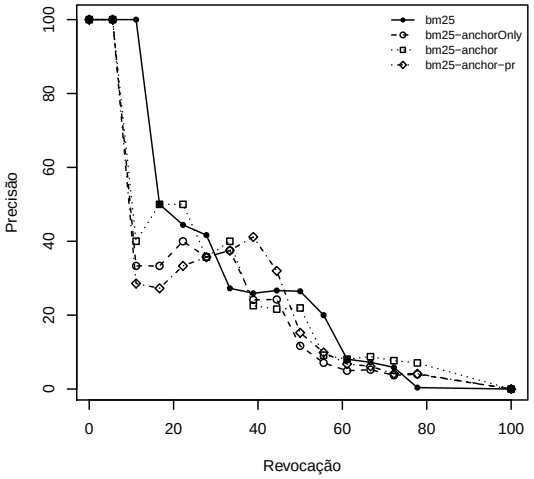

funk
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

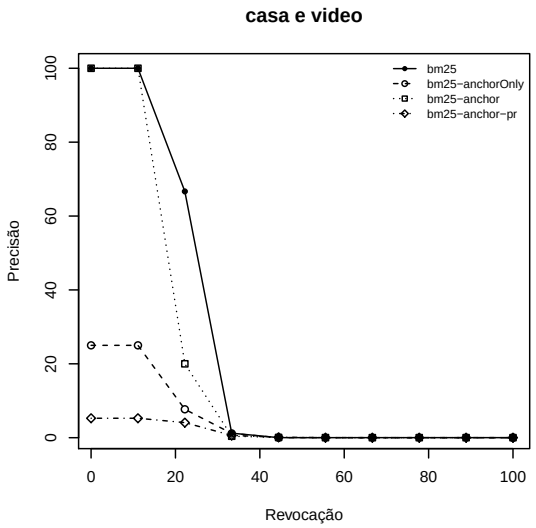

caixa economica federal
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

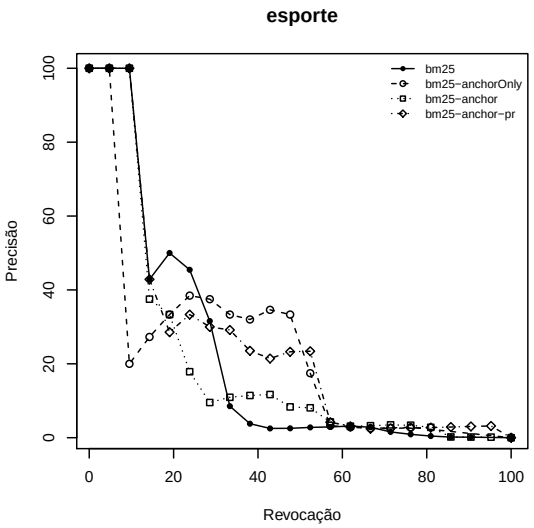

mercado livre
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

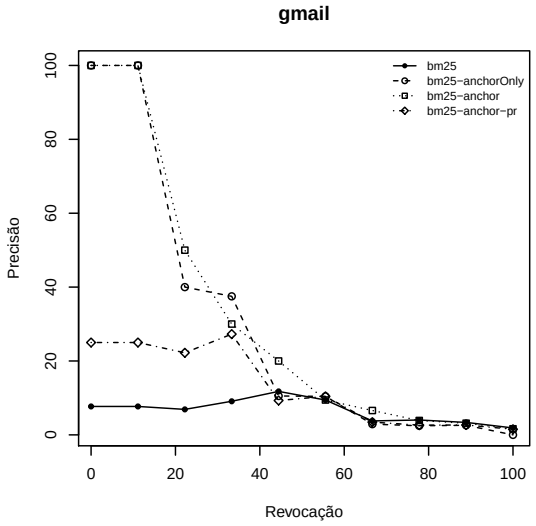

youtube
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

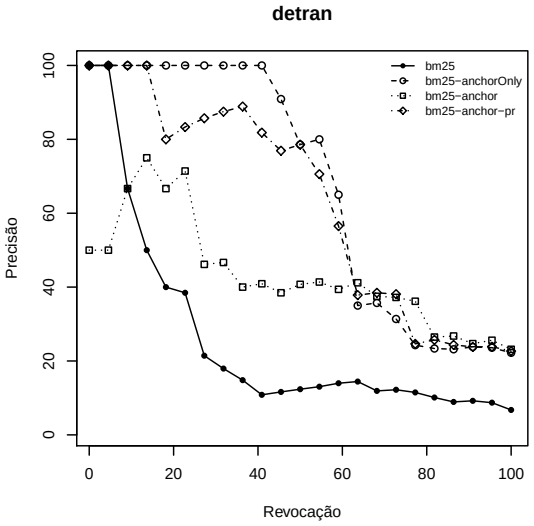

naruto
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

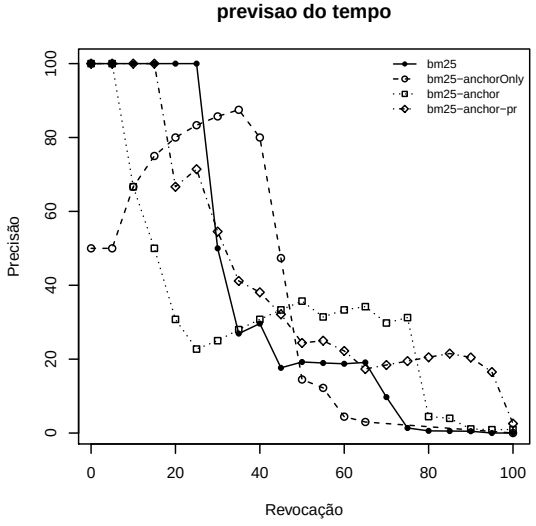

oi
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

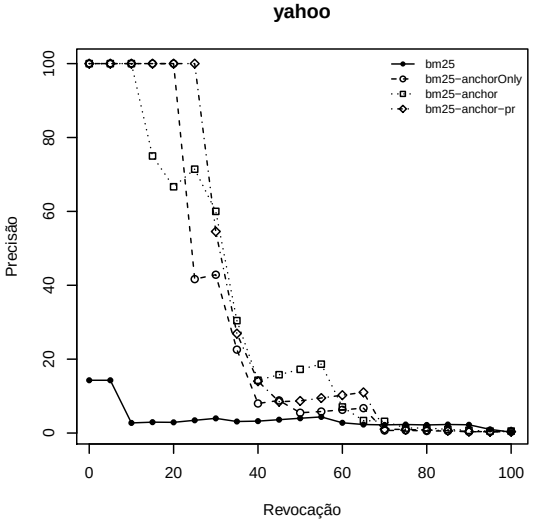

casa e video
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

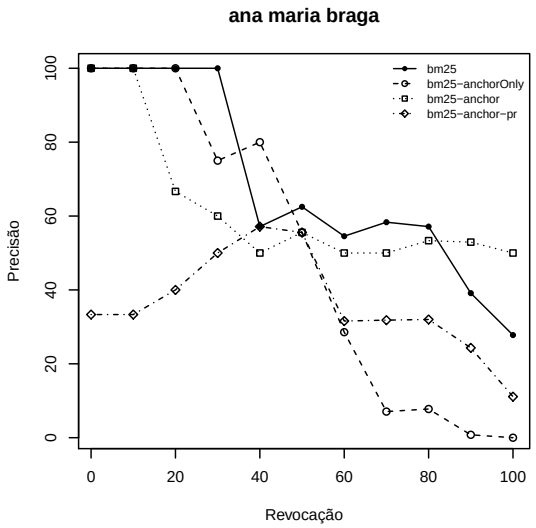

esporte
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

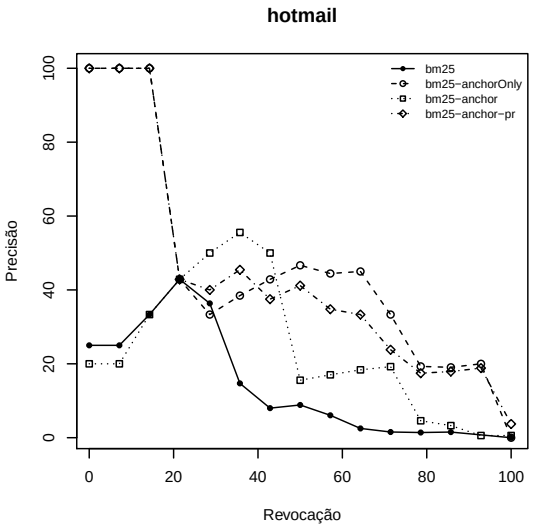

gmail
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

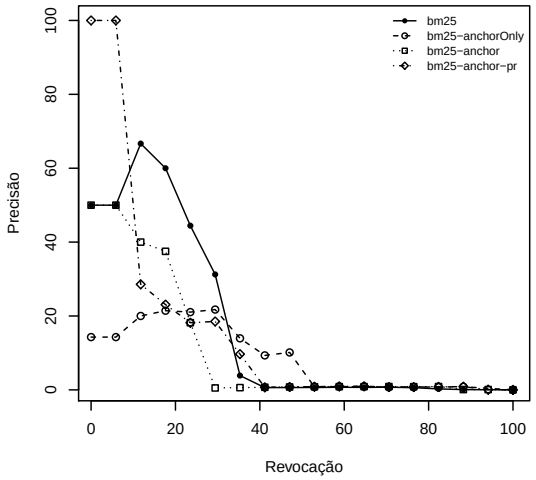

detran
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100

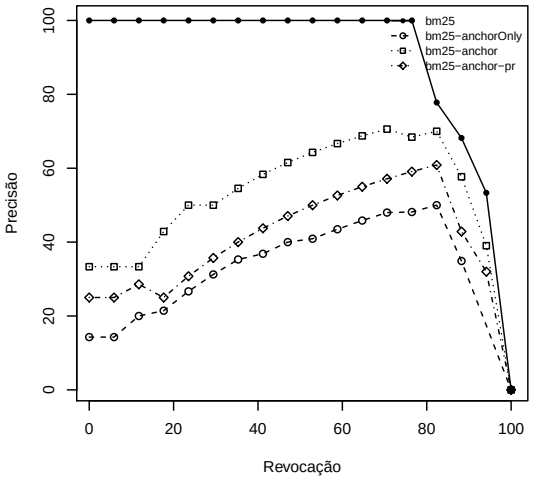

previsao do tempo
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100


yahoo
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
Precisão
Revocação
0
20
40
60
80
100


ana maria braga
Precisão
Revocação
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
0
20
40
60
80
100


hotmail
Precisão
Revocação
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
0
20
40
60
80
100


msn
Precisão
Revocação
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
0
20
40
60
80
100


panico
Precisão
Revocação
bm25
bm25−anchorOnly
bm25−anchor
bm25−anchor−pr
0
20
40
60
80
100

4. Conclusão

Neste trabalho foi possível verificar que a introdução de mais fontes de evidência são importantes para se obter melhores resultados. No entanto, da mesma forma que as informações ajudam, elas podem introduzir ruídos que prejudiquem as consultas.

Enquanto o texto das âncoras se mostrou quase sempre benéfico, o *Pagerank* se mostrou mais direcionado para determinadas situações onde a consulta é navegacional. Desta forma, maiores estudos seriam necessários para entender melhor a influência do mesmo.