RI – Relatório do Trabalho Prático 2

Danilo Ferreira e Silva[1]

[1]Departamento de Ciência da Computação – UFMG

danilofs@dcc.ufmg.br

1. Introdução

O objetivo deste trabalho foi a implementação de um processador de consultas baseado no modelo vetorial, BM25 e uma combinação dos dois. Além disso, o sistema desenvolvido foi avaliado com base em um conjunto de consultas de referência, cada uma contendo o respectivo conjunto de documentos relevantes.

O sistema está disponível para avaliação em:

`http://scala.llp2.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 algoritmo implementado para recuperar os documentos é o muito semelhante ao algoritmo base visto em aula. Os termos da consulta são ordenados pelo inverso de suas frequências na coleção e processados nessa ordem.

É usado um número máximo de acumuladores configurável e política *continue*. Ou seja, mesmo após o limite de acumuladores ser excedido as listas invertidas continuam sendo lidas e aqueles documentos que já possuem acumulador recebem a contribuição do termo.

Esta etapa tem custo linear com relação a soma dos tamanhos das listas invertidas de cada termo. Por esse motivo, para consultas com muitos termos, ou termos que aparecem em muitos documentos, o tempo de processamento pode ser muito alto.

Após o cálculo da contribuição dos termos. Parte-se para a ordenação dos r documentos do topo, o que é feito com um *heap* de tamanho r. O *heap* é preenchido com r elementos e contruído. Em seguida são feitas no máximo $a - r$ operações de reconstruir o *heap*, onde a é o número de acumuladores usados na consulta. Logo, o custo desta etapa é $O(r + alog(r))$.

2.1. Modelo vetorial

O modelo vetorial foi implementado com base no seguinte esquema de pesos:

$$sim_v(Q, D) = \frac{1}{W_d} \sum_{k_i} (1 + log(f_d, t)) log(1 + N/n_t)$$

A norma W_d de cada documento foi definida como:

$$W_d = \sqrt{\sum_{k_i}(1 + log(f_{d,t}))^2}$$

Estes valores são pré-cumputados em tempo de indexação, e armazenados em um arquivo a parte.

2.2. Modelo BM25

O modelo implementado foi baseado na seguinte equação:

$$sim_b(Q, D) = \sum_{k_i}(\frac{(K_1 + 1)f_{i,j}}{K_1((1 - b) + b(len_d/avglen)) + f_{i,j}})(\frac{N - n_t + 0.5}{n_t + 0.5})$$

Para este modelo len_d foi considerado a quantidade de palavras totais do documento. $avglen$ é a média de len_d na coleção. Estes valores também foram pré-computados em tempo de indexação.

2.3. Modelo combinado

O modelo combinado utilizado é uma combinação linear dos dois modelos acima, controlados por uma constante $0 <= C <= 1$.

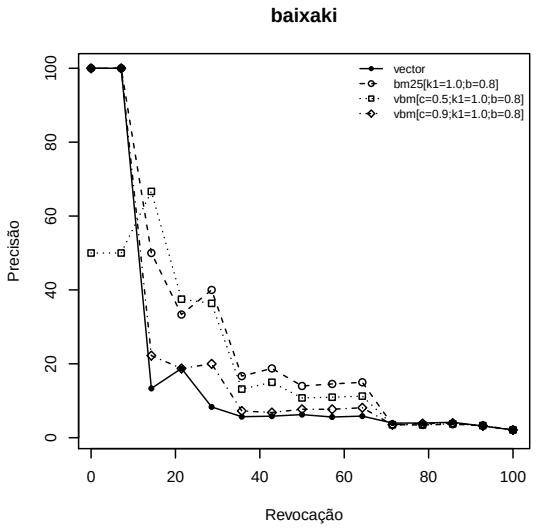
$$sim_{v,b}(Q, D) = Csim_v, b(Q, D) + (1 - C)sim_v, b(Q, D)$$

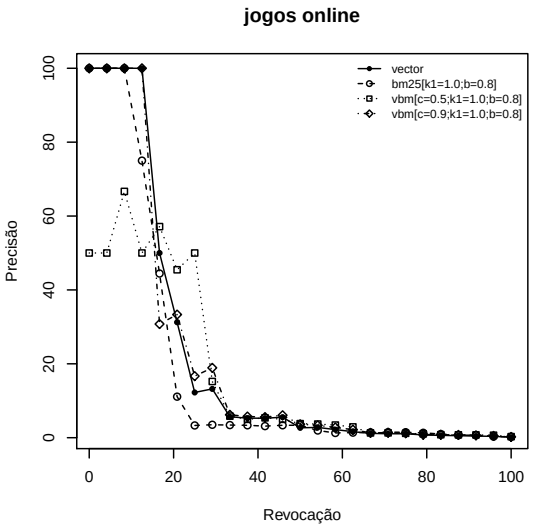
3. Avaliação

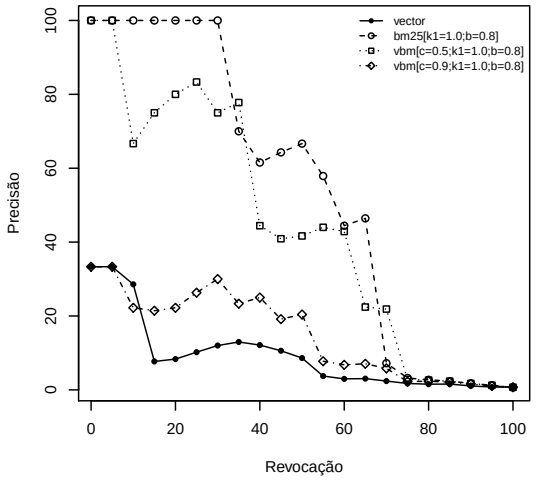
Foram executadas todas as consultas de referência nos três modelos implementados e geradas curvas de precisão por revocação, conforme figuras a seguir.

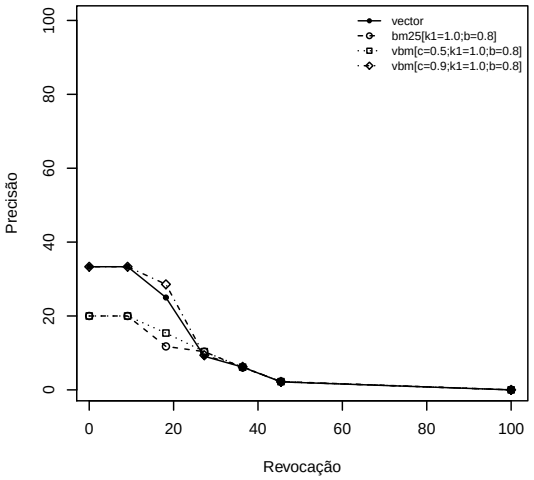
Para traçar estes gráficos, o número de resultados e acumuladores máximos foi definido como ilimitados, para que se chegasse a 100% de *recall*. Um fato observado foi que nem toda consulta foi possível chegar a 100% *recall*. Uma explicação para este fato é que nem todo documento relevante para uma consulta tem de fato termos da mesma em seu conteúdo. Um exemplo de tal documento é o `http://clubedepoquer.com/`, que não contém o termo poquer.

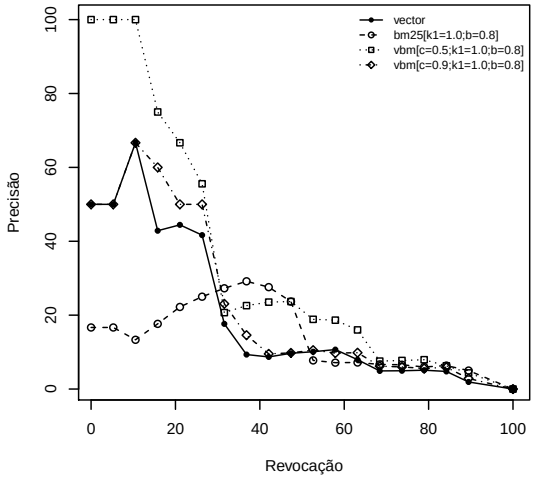
Não foi encontrado um modelo que fosse consistentemente melhor do que o outro, embora o BM25 tenha sido melhor na maioria das consultas. Para o modelo combinado, não foi possível encontrar um valor de C que fosse consistemente melhor que o próprio BM25 puro. Nos gráficos são utilizados os valores 0.5 e 0.9.

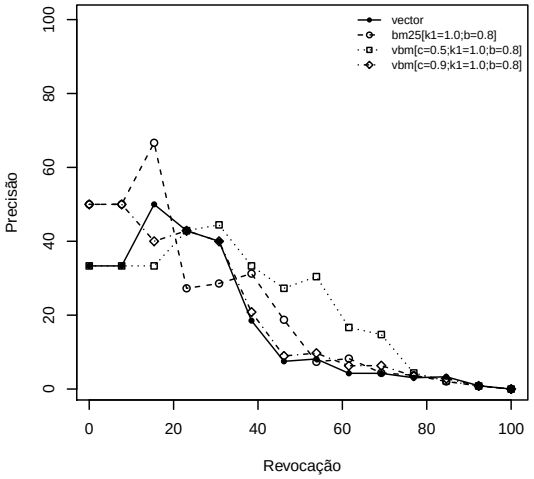

baixaki
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

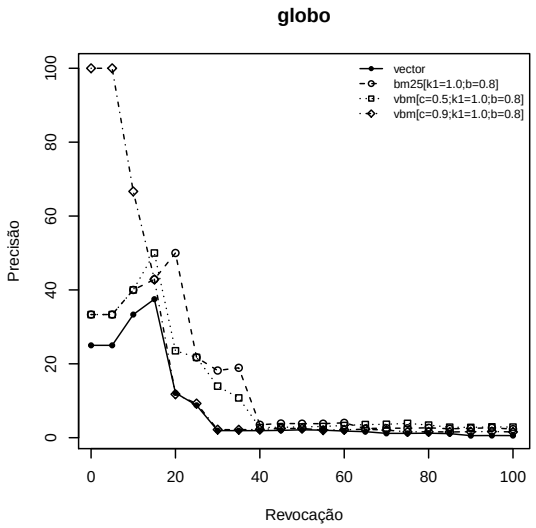

jogos online
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

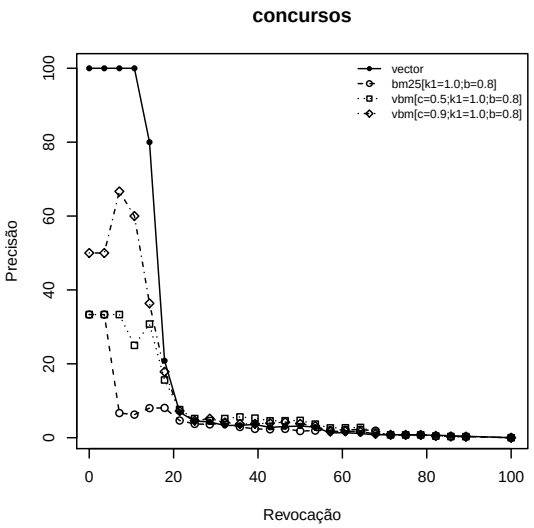

receita federal
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

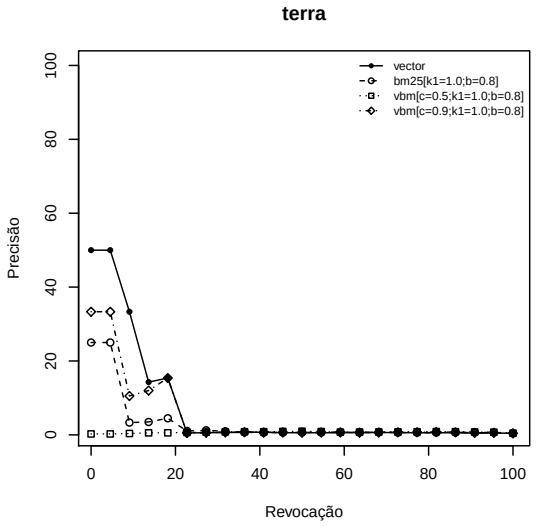

poquer
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

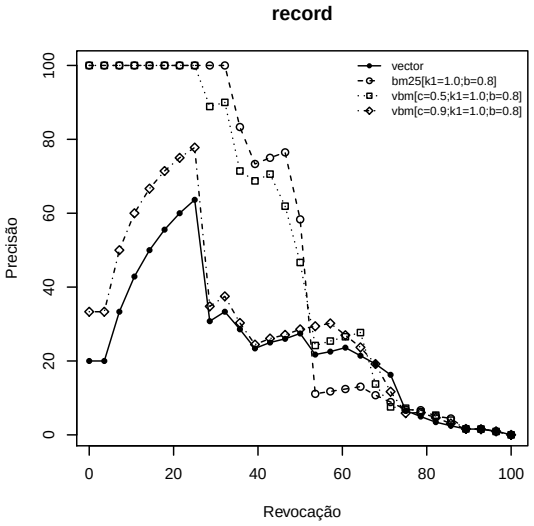

ig
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

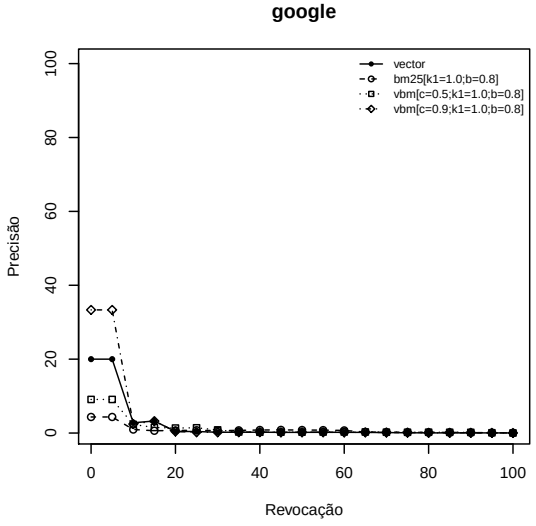

claro
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

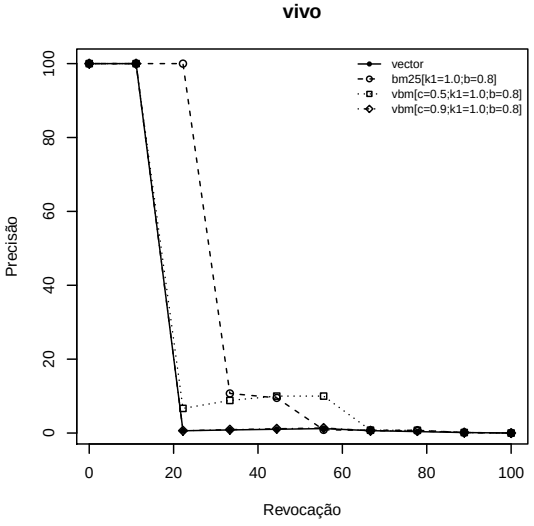

globo
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

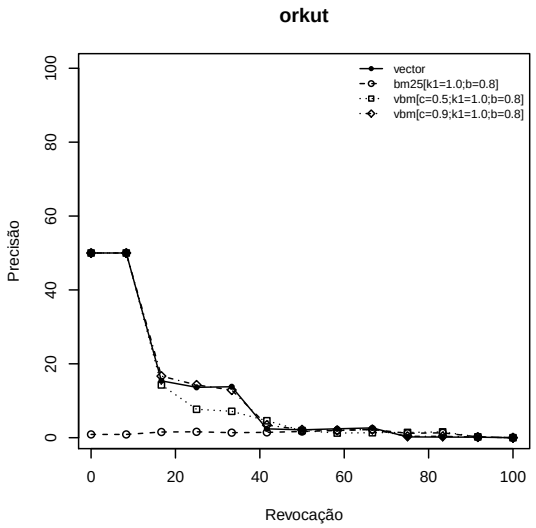

concursos
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

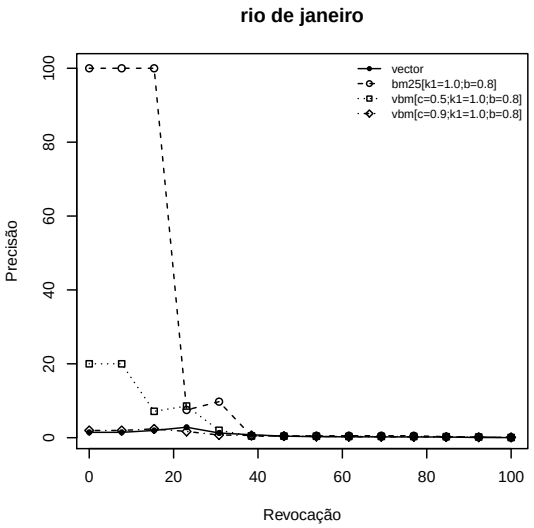

terra
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

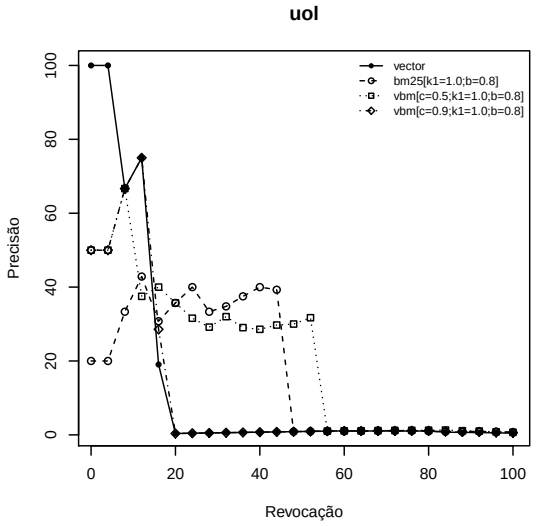

record
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

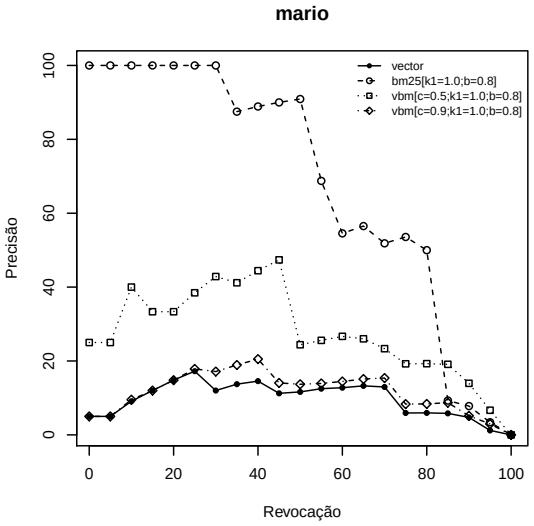

google
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

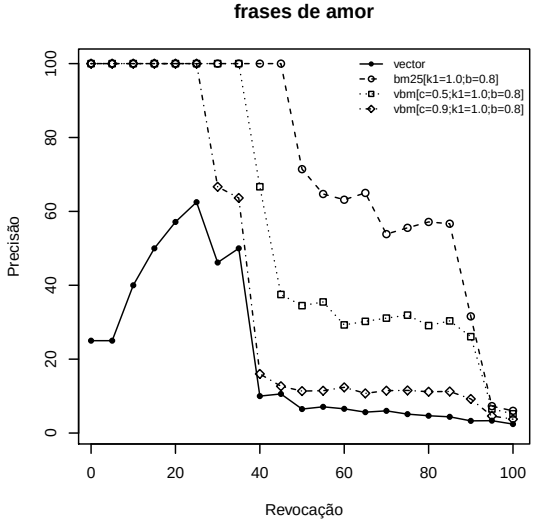

vivo
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

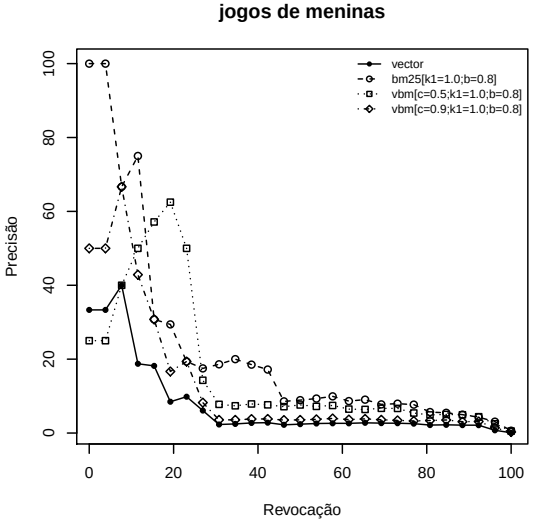

orkut
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

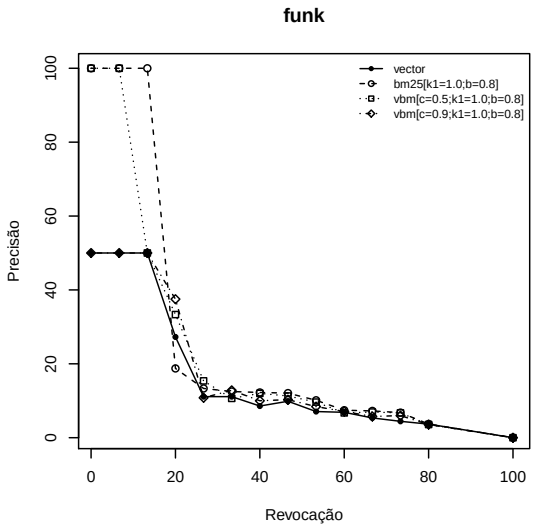

rio de janeiro
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

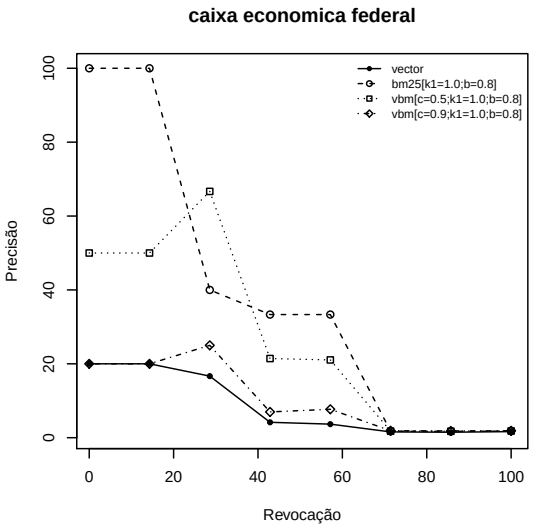

uol
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

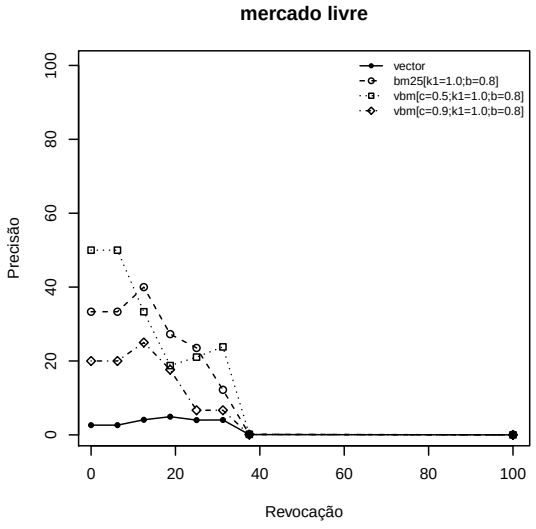

mario
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

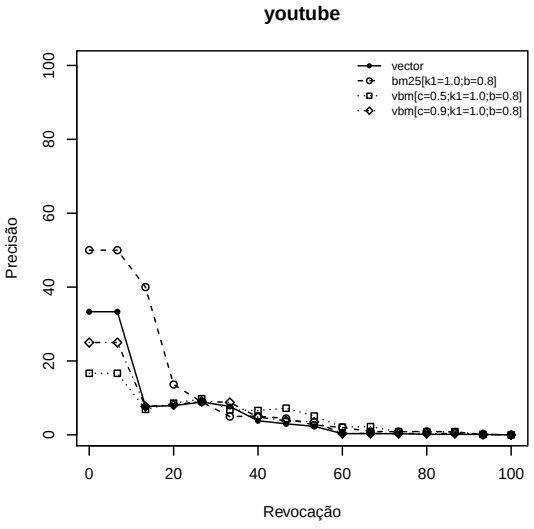

frases de amor
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

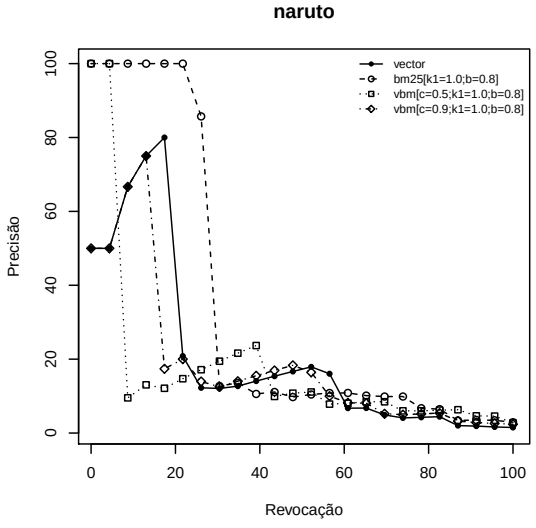

jogos de meninas
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

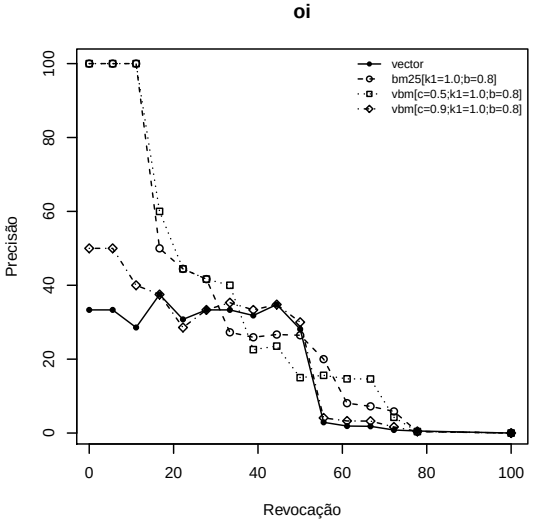

funk
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

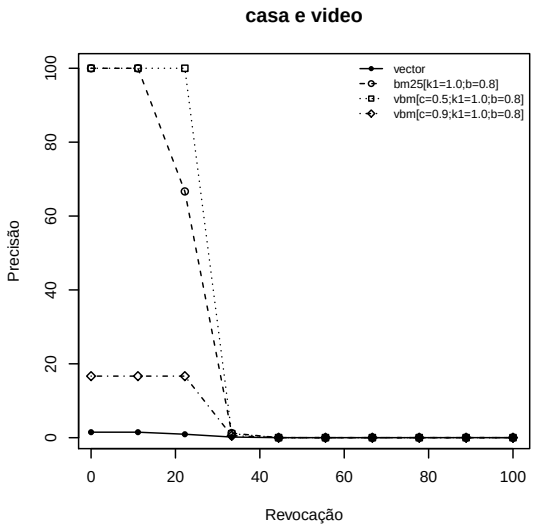

caixa economica federal
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

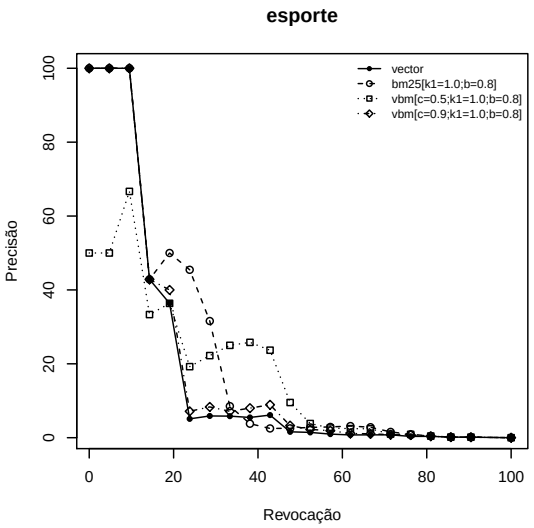

mercado livre
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

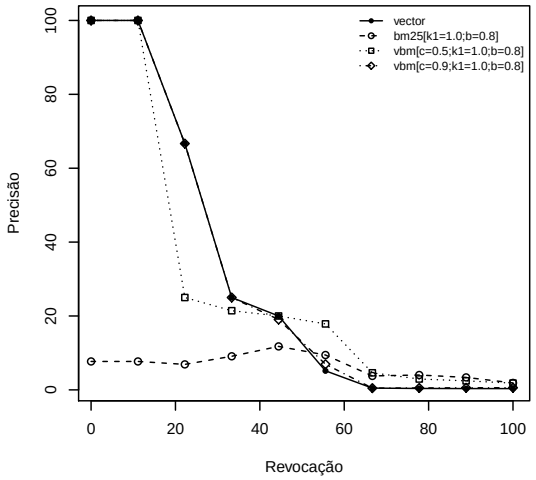

youtube
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

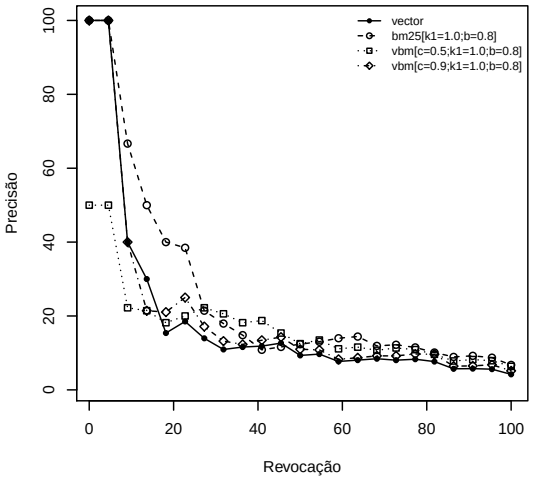

naruto
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

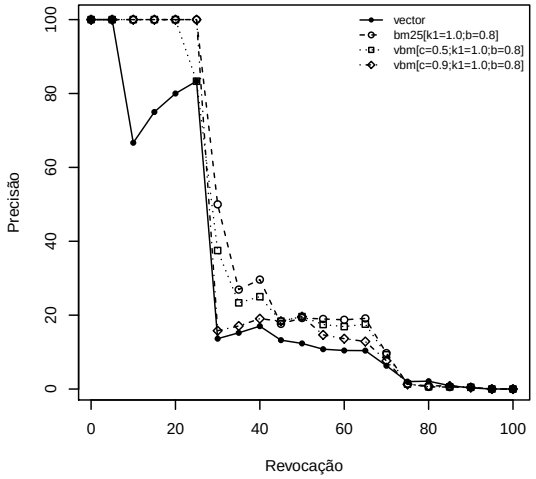

oi
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

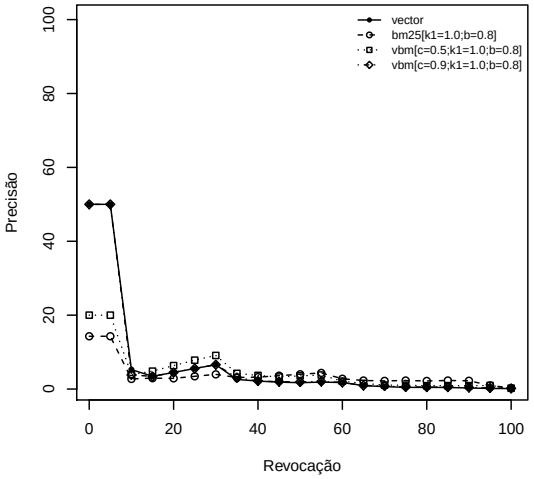

casa e video
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

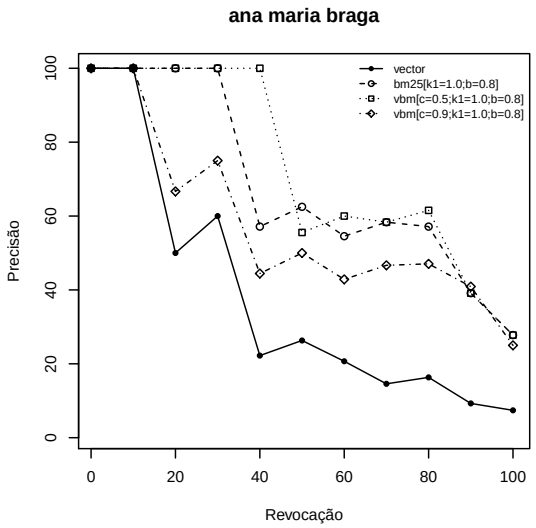

esporte
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

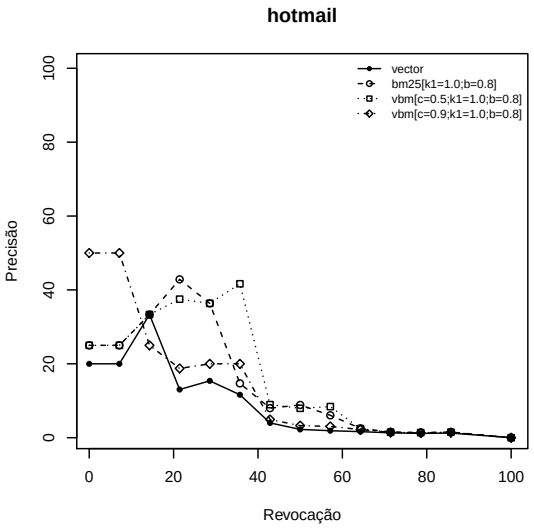

gmail
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

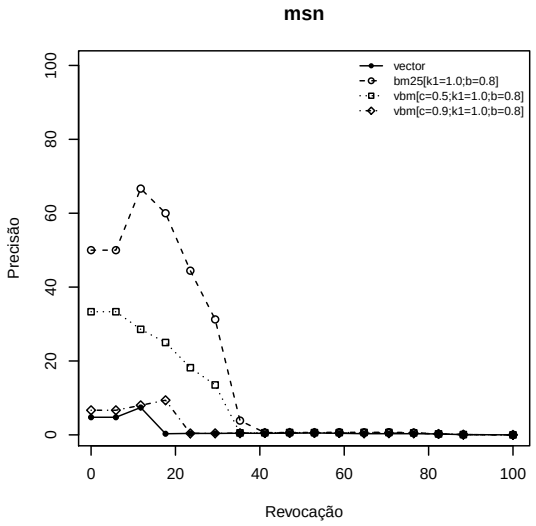

detran
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100

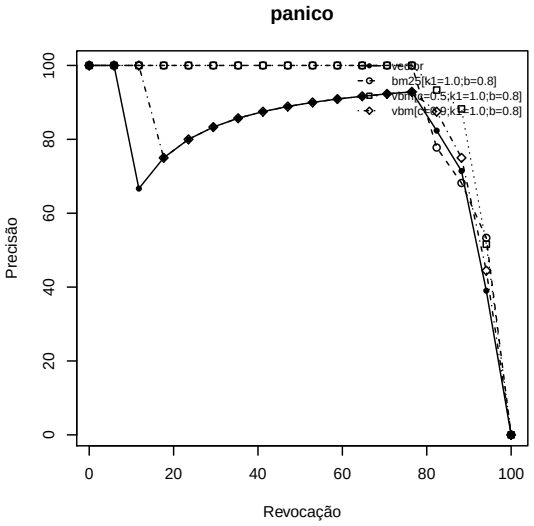

previsao do tempo
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100


yahoo
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]
Precisão
Revocação
0
20
40
60
80
100


ana maria braga
Precisão
Revocação
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]


hotmail
Precisão
Revocação
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]


msn
Precisão
Revocação
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]


panico
Precisão
Revocação
vector
bm25[k1=1.0;b=0.8]
vbm[c=0.5;k1=1.0;b=0.8]
vbm[c=0.9;k1=1.0;b=0.8]

4. Conclusão

Com este trabalho foi possível observar que o problema de *ranking* dos documentos é complexo e apenas os termos presentes no mesmo não traz resultados satisfatórios. Isto porque a diferenciação entre documentos que contém os mesmos termos é muito pobre, pois apenas as frequências dos termos nos documentos diz muito pouco.

Outras fontes de evidência devem ser exploradas para que se tenha resultados melhores, tais como estrutura do documento, texto de âncoras e *page rank*.