Recuperação de Informação e Melhoria de Resultados com o uso de Sintagmas Nominais

Alair Pereira do Lago
Wendel Scardua

09 de novembro de 2010

Visão Geral

Sistemas de Recuperação de Informação (R.I.)

- Em geral, permitem buscas exatas
- Qualidade dos resultados é medida através da precisão e da cobertura
  - **Precisão** - dos documentos retornados, quantos são relevantes
  - **Cobertura** - dos documentos relevantes, quantos foram retornados

Problemas

- **Sinonímia** - muitas expressões para um mesmo conceito
  - "Pé de laranja", "Árvore de laranja", "Laranjeira"
  - baixa cobertura
- **Polissemia** - muitos conceitos para uma mesma expressão
  - "Pé (membro)", "Pé (árvore)", "Pé (unidade de medida)"
  - baixa precisão

Abordagens

- Expansão de consulta
- Uso de ontologias
- Processamento de Linguagem Natural
- Indexação Semântica Latente (Latent Semantic Indexing, ou LSI)
- Grafos de Rede Contextual (Contextual Network Graphs)

Expansão de Consulta

- Modifica a consulta original, para aumentar a cobertura
- Porém, pode diminuir a precisão
- Divide-se em:
  - **Análise Global** - trabalha sobre a coleção de documentos para obter termos novos
  - **Análise Local** - trabalha sobre resultados da consulta original

  e pode ser:
  - **Manual** - o usuário interage com o processo de expansão de consulta
  - **Automática** - a expansão é feita sem intervenção do usuário

Ontologias

- Representações de conceitos
- Conjuntos de sinônimos, hierarquia de hiperônimos (*WordNet*)
  - “carro” é um “veículo”
  - “pé” é um “membro”
  - “pé” é uma “unidade”

Expansão de Consulta via Ontologias

- Adiciona-se à consulta original sinônimos e hiperônimos, obtidos da ontologia.
  - Exemplo: numa consulta por “remédio”, adiciona-se “medicamento”, caso na ontologia conste essa relação
- Pode-se atribuir pesos diferentes aos termos adicionados, dependendo da relação com os termos originais

Processamento de Linguagem Natural

Alguns recursos de P.L.N. que podem ser utilizados em Recuperação de Informação:

- Lematização (ou stemming)
- Identificação de sintagmas nominais

Lematização

- Reduz um termo ao seu lema. Por exemplo:
  - *professoras* $\leftarrow$ *professor*
  - *latinha* $\leftarrow$ *lata*
- Com isso atenua-se o problema da sinonímia
- Geralmente utiliza-se "stemming" como aproximação; consiste em se aplicar um conjunto de regras a fim de reduzir uma palavra a um radical, mas apenas aproximadamente
  - Exemplo: remover o final "-s" para transformar plural em singular

Sintagmas Nominais

- Termos isolados podem sofrer polissemia
- Sintagmas nominais fornecem contexto, por exemplo:
  - “pé de laranja”
  - “pé da mesa”
- Pode-se fazer expansão de consulta através de sintagmas

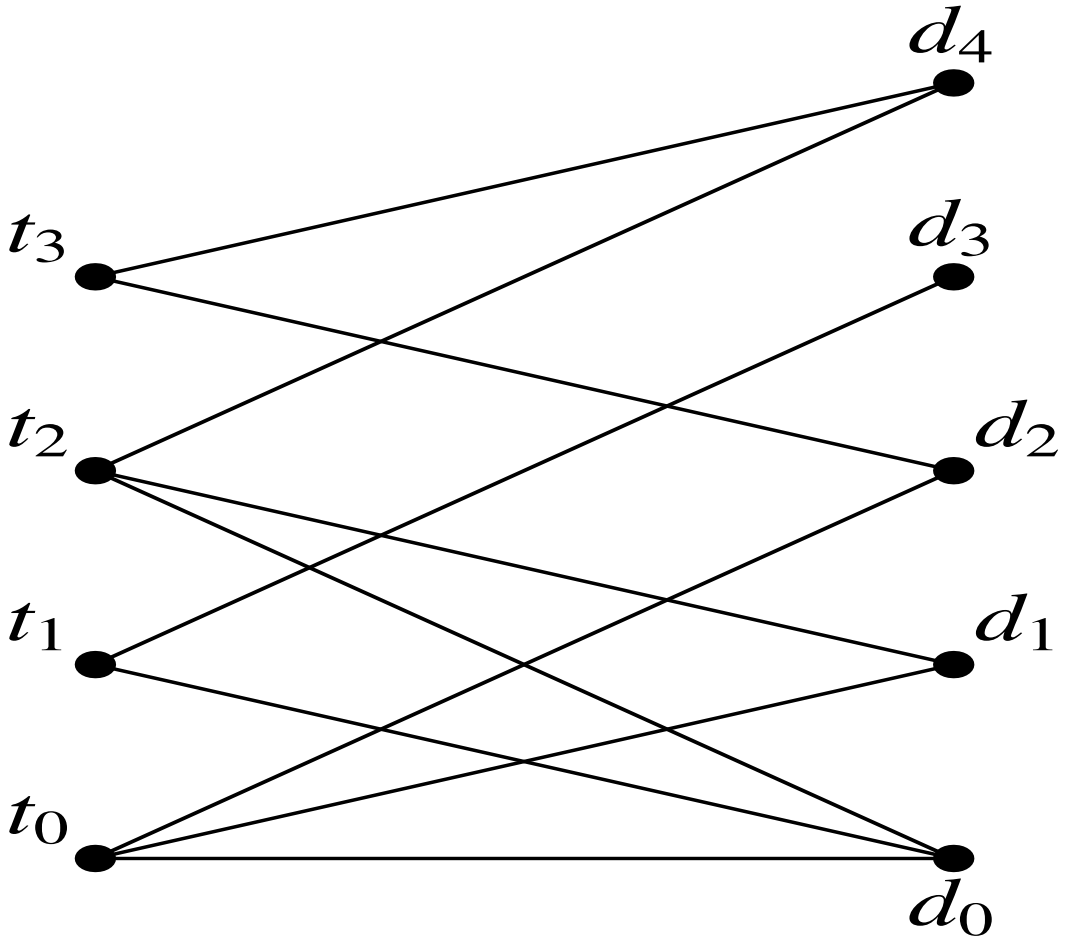
Indexação Semântica Latente (LSI)

- Faz uso de uma matriz termos-por-documentos
- Consiste em utilizar recursos de Álgebra Linear para analisar a relação entre termos e documentos através dessa matriz
- Geralmente envolve Decomposição em Valores Singulares (SVD)
- Experimentos não foram feitos devido ao custo computacional do SVD

Grafos de Rede Contextual

- Estrutura consiste em um grafo bipartido, em que os vértices são termos e documentos.
- Arestas conectam documentos a seus respectivos termos, e o peso de uma aresta corresponde ao peso do termo no documento (como o tf.idf)
- Para realizar uma consulta:
  - Cada vértice começa com "energia" zero;
  - Para cada termo da consulta, seu respectivo vértice recebe uma energia E_0
  - Cada vértice que recebe uma energia superior a um limite T, a divide para seus vizinhos
  - Ao percorrer uma aresta, essa parte da energia decai em função do peso
  - Ao final do processo, os documentos com maior energia podem ser retornados, e a energia mede sua relevância

Grafos de Rede Contextual



Grafos de Rede Contextual

- Como no LSI, permite encontrar documentos relacionados, mesmo sem os termos originais
- Para um termo qualquer, quanto mais documentos ele possuir em comum com termos da consulta, maior será a energia resultante nele - isso permite descobrir termos relacionados.
- Dado que o grafo não faz distinção entre documentos e termos, é possível realizar consultas a partir de documentos a fim de encontrar outros similares.

Experimentos

- Metodologia e métricas
- Sintagmas Nominais
- Grafos de Rede Contextual
- Conclusão

Metodologia de Avaliação dos Resultados

- Metodologia e métricas utilizadas pelo CLEF - Cross-Language Evaluation Forum, na trilha "Monolingual Ad-hoc" de 2006 para língua portuguesa
- Coleção: notícias em português, dos jornais Folha de São Paulo (Brasil) e Público (Portugal), entre 1994 e 1995
- Teste: 50 tópicos; cada um possuindo:
  - um título, com poucas palavras
  - uma descrição mais detalhada do tópico
  - uma lista de dezenas de documentos da coleção, classificados como sendo relevantes ou não para o tópico

Métricas

Seja q uma consulta, e seja $D(q)$ o seu respectivo conjunto-resposta, ordenado segundo um índice de relevância.
Definimos $D_r(q)$ como sendo o subconjunto de $D(q)$ formado por seus primeiros r documentos.
Seja $R(q)$ o conjunto de todos os documentos realmente relevantes para a consulta q.
A *cobertura* (*recall*) é dada por:

$$\rho_r(q) = \frac{|R(q) \cap D_r(q)|}{|R(q)|} \tag{1}$$

E a *precisão* (*precision*) é dada por:

$$\pi_r(q) = \frac{|R(q) \cap D_r(q)|}{|D_r(q)|} \tag{2}$$

- r baixo - o usuário tem interesse em poucos documentos, mas que sejam altamente relevantes
- r alto - o usuário quer obter o maior número possível de documentos relevantes

Exemplo de Cobertura e Precisão

posição	relevante ?	$\rho_r(q)$	$\pi_r(q)$
1	sim	0.20	1.00
2	não	0.20	0.50
3	não	0.20	0.33
4	sim	0.40	0.50
5	não	0.40	0.40
6	não	0.40	0.33
7	sim	0.60	0.43
8	sim	0.80	0.50
9	não	0.80	0.44
10	sim	1.00	0.50

Métricas derivadas de cobertura e precisão

- Percorrendo-se os valores possíveis de r, é possível traçar um gráfico de precisão por cobertura.
- Normalmente se faz uma interpolação desse gráfico: para valor possível de cobertura $\rho \in [0, 1]$, atribuimos um valor de precisão, dado por:

$$\Pi_q(\rho) = \max_r \left\{ \pi_r(q) | \rho_r(q) \geq \rho \right\} \tag{3}$$

- Assim se obtém uma curva monotonicamente decrescente, tal que cada valor de cobertura está associado a apenas um valor de precisão.
- Para se obter um gráfico independente da consulta realizada, toma-se a média de todas as consultas:

$$\Pi(\rho) = \frac{1}{|Q|} \sum_{q \in Q} \Pi_q(\rho) \tag{4}$$

Exemplo

posição	relevante ?	$\rho_r(q)$	$\pi_r(q)$	$\Pi_q(\rho)$
1	sim	0.20	1.00	1.00
2	não	0.20	0.50	1.00
3	não	0.20	0.33	1.00
4	sim	0.40	0.50	0.50
5	não	0.40	0.40	0.50
6	não	0.40	0.33	0.50
7	sim	0.60	0.43	0.50
8	sim	0.80	0.50	0.50
9	não	0.80	0.44	0.50
10	sim	1.00	0.50	0.50

Gráfico

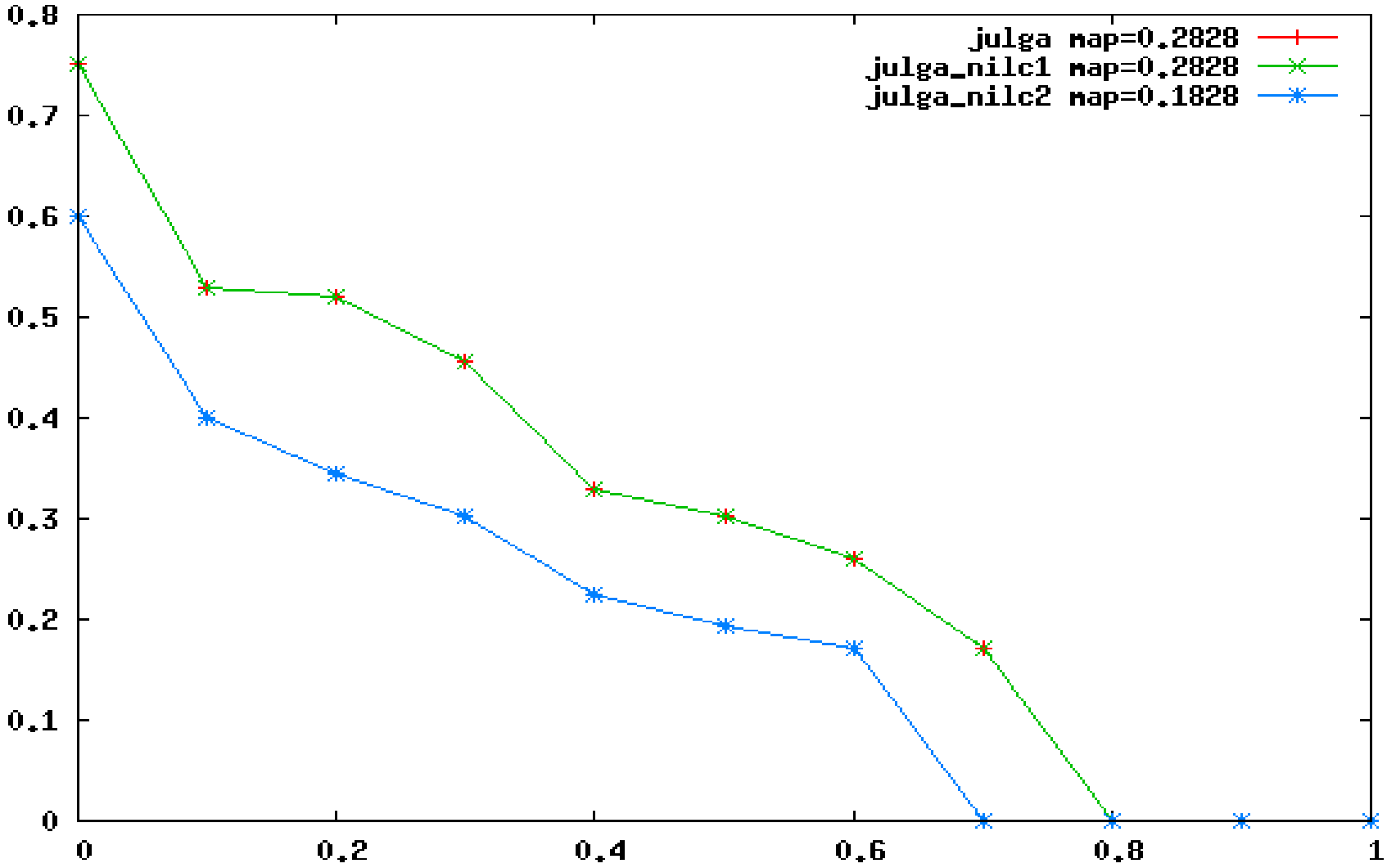


Figura: Gráfico de Precisão (Y) por Cobertura (X)

Precisão média

Dado o conjunto $D_r(q)$, a precisão média consiste na média dos valores $\pi_j(q)$ para cada j tal que $d_j \in D_r(q)$ seja de fato relevante.
No exemplo anterior, a precisão média é a média das precisões relativas às posições 1, 4, 7, 8 e 10, ou seja,
$\frac{1}{5}(1.00 + 0.50 + 0.43 + 0.50 + 0.50) = 0.586$
Chamamos de *MAP* (*mean average precision*) a média dessas precisões médias, tomada ao longo de uma ou mais consultas q.

Experimentos com Sintagmas Nominais

- Selecionamos os K primeiros documentos
- Para cada ocorrência, tomamos x sentenças ao redor (análise local)
- Coletamos todos os sintagmas nominais destas sentenças
- Contabilizamos a frequência de cada termo na coleção de sintagmas
- Tomamos como o peso de cada sintagma a soma das frequências de seus núcleos
- Adicionamos os s sintagmas de maior peso à consulta original

Exemplo

Para a consulta: `+desempreg* +europ*`, um trecho de documento devolvido seria:

```
{[JCP]} lança {[campanha] contra o [desemprego]} .
"Sem {[emprego]} nada feito" é {o [lema] de a [campanha]}
que {a [Juventude Comunista Portuguesa]} ( {[JCP]} ) vai
lançar ainda {este [ano]} e que pretende discutir e avançar
{[propostas]} para {o [problema] de o [desemprego] de
os [jovens]} .
```

Sintagmas Obtidos

Sintagma Nominal	Sub-consulta
JCP	(+JCP)
campanha contra o **desemprego**	(+campanha +desemprego)
emprego	(+emprego)
o **lema** de a **campanha**	(+lema +campanha)
a **Juventude Comunista Portuguesa**	(+Juventude +Comunista +Portuguesa)
este **ano**	(+ano)
propostas	(+propostas)
o **problema** de o **desemprego** de os **jovens**	(+problema +desemprego +jovens)

Experimentos

Parâmetros:

- K – quantos documentos considerados relevantes serão analisados
  $K = 10, 20, 40, 80, 160, 320, 5\%, 10\%, 20\%, 40\%, 100\%$
- x – até quantas sentenças farão parte de cada passagem
  $x = 1, 3, 5, 9, 17, 33$ (0, 1, 2, 4, 8 e 16 linhas antes e depois)
- s – quantos sintagmas serão usados
  $s = 15, 30, 45, 60, 120, 25\%, 50\%, 100\%$
- Total: 528 combinações

De todas as combinações:

- apenas 15 obtiveram MAP menor do que o padrão 34,01%
- a melhor foi ($K = 80, x = 1, s = 100\%$), MAP é de 44,56%, 31% acima do padrão

Experimentos

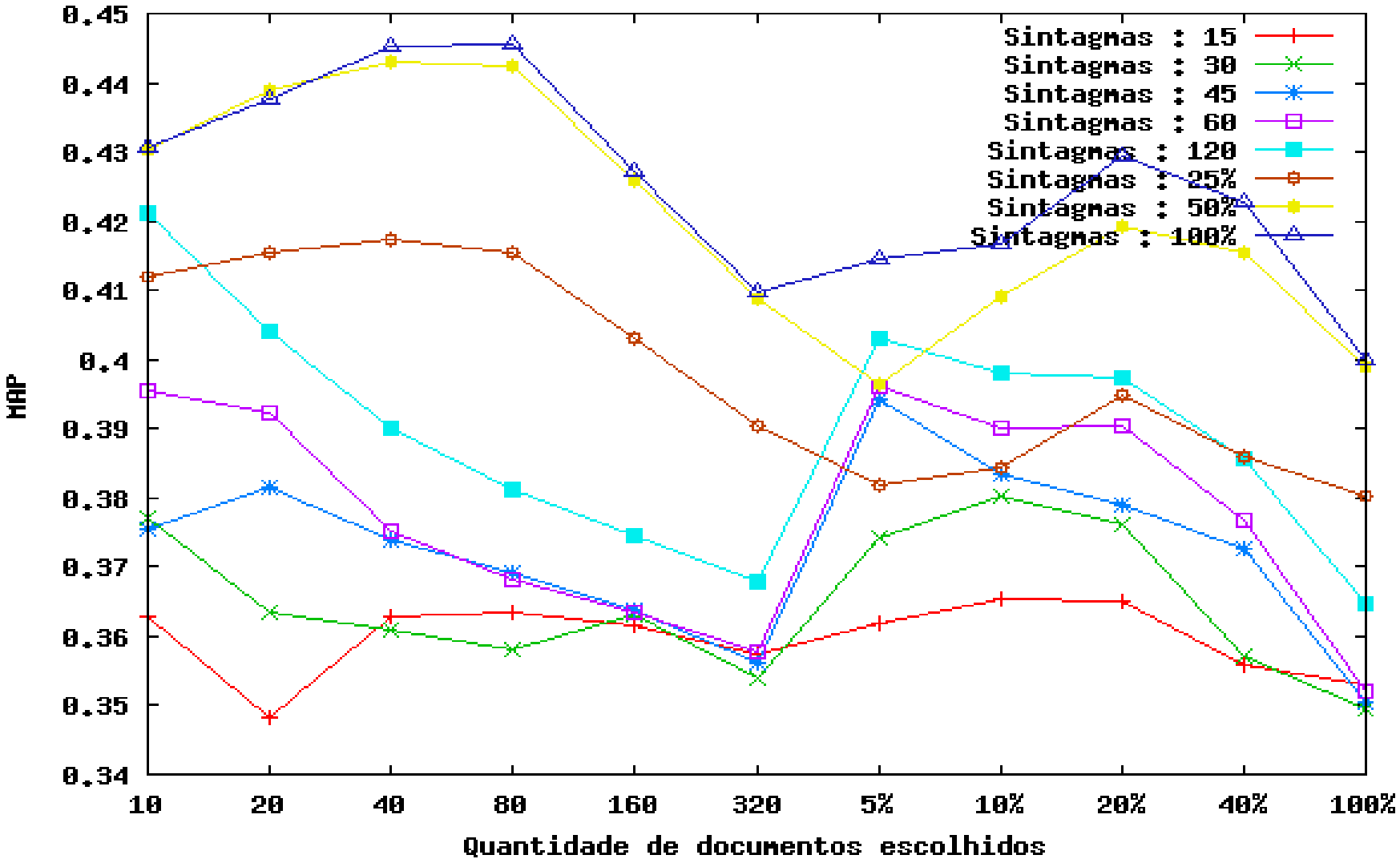


Figura: Valores de MAP para variações de s e K, fixando $x = 1$

Experimentos

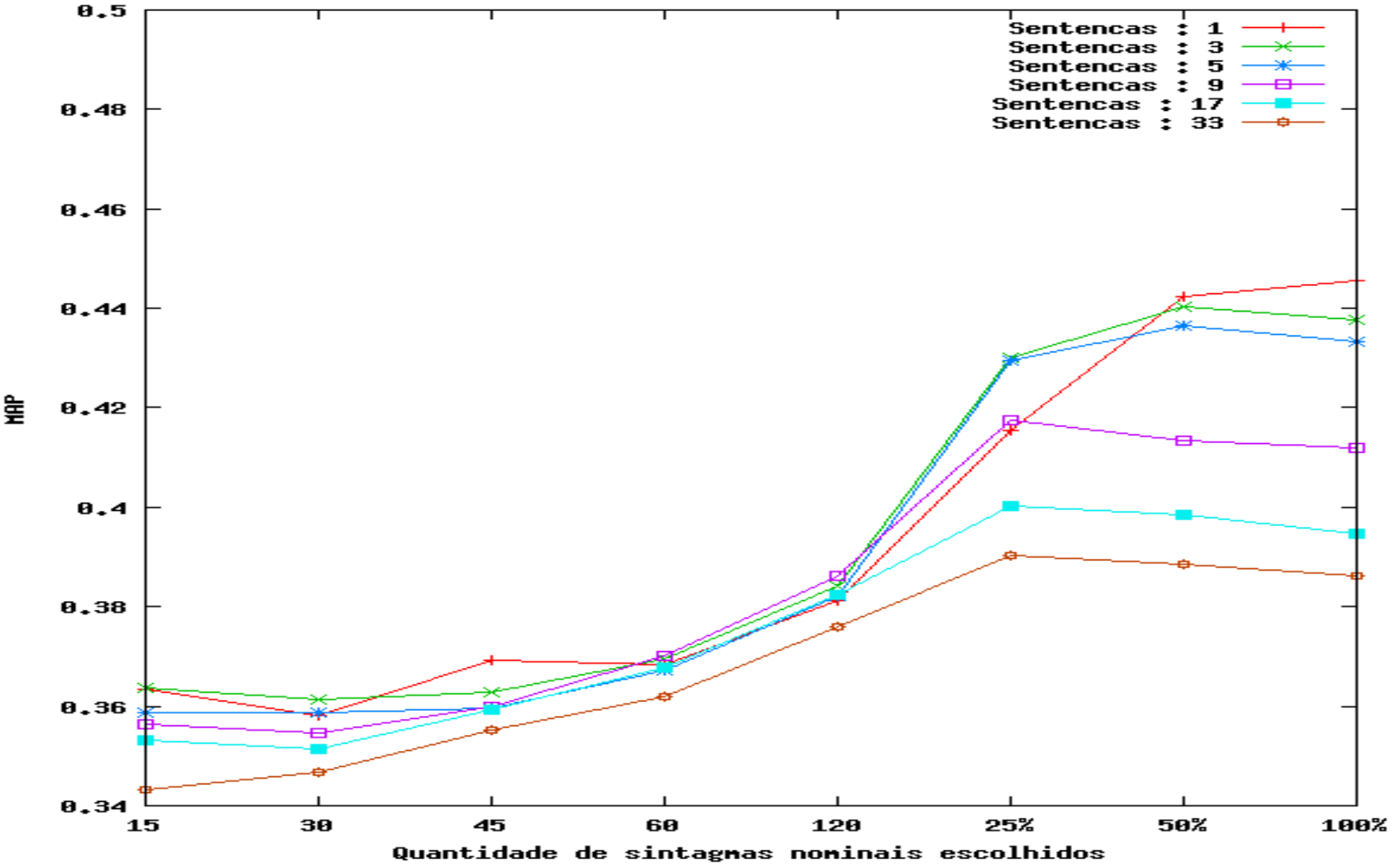


Figura: Valores de MAP para variações de x e s, fixando $K = 80$

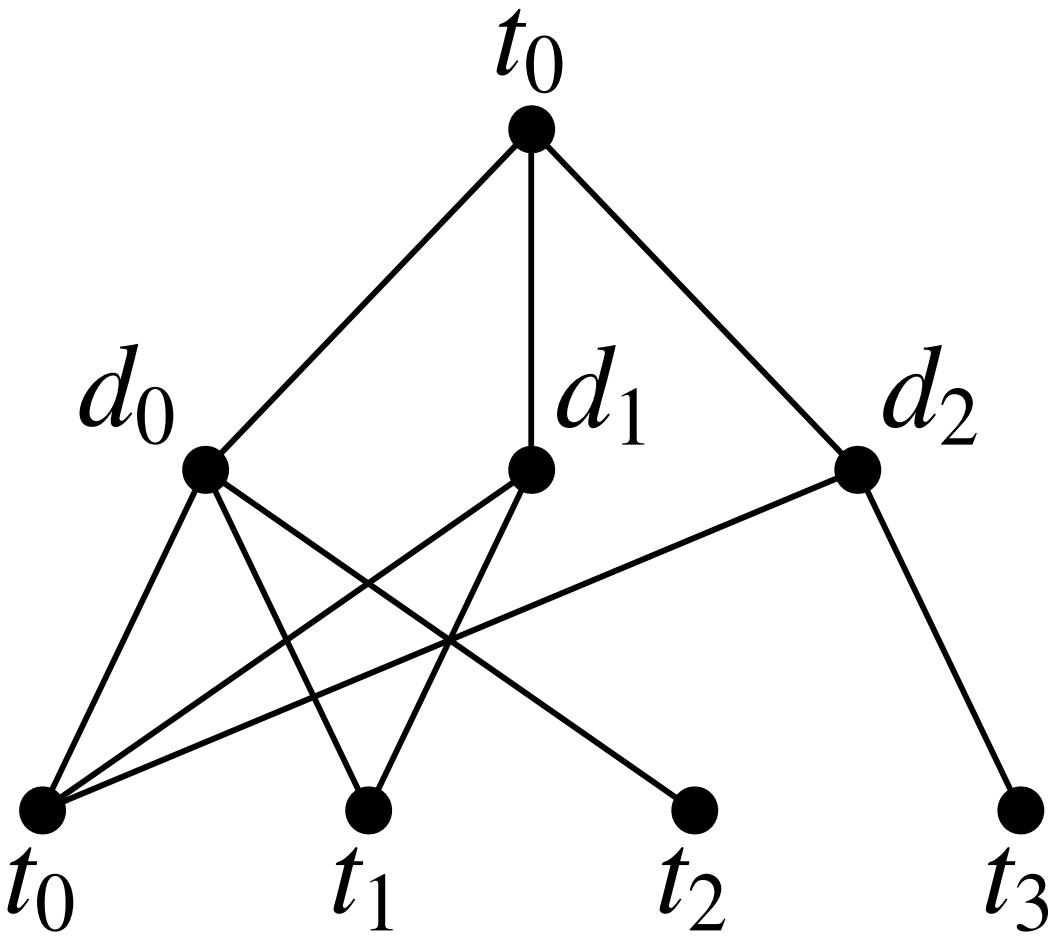
Grafos de Rede Contextual

- Algoritmo próprio, inspirado no artigo original
- Original não descrevia valores de parâmetros e resultados
  - nossa tentativa de implementá-lo resultou num MAP de 32,50%
- Duas estratégias foram criadas:
  - Documentos-para-Termos, ou $D \rightarrow T$ – pela facilidade de implementação
  - Termos-para-Termos, ou $T \rightarrow T$ – para ser mais fiel ao original

Estratégia $T \rightarrow T$

- Obtém-se os termos da consulta original
- Aplica-se o algoritmo de busca CNG sobre os vértices desses termos
- Gera-se uma nova consulta a partir dos pares (termo, energia) obtidos
- Realiza-se uma nova pesquisa com esta consulta

Estratégia $T \to T$



Estratégia $T \rightarrow T$

- Referentes a busca CNG:
  - E_0 – energia inicial
  - t – *threshold*
  - f – taxa de transmissão
  - $\delta(x)$ – função aplicada ao peso da aresta
- Referentes a geração da consulta:
  - c – booleano, define se a consulta original será adicionada
  - γ – fator que divide a energia final dos termos
  - $\lambda(x)$ – função aplicada à energia resultante

Estratégia $T \rightarrow T$

- Referentes a busca CNG:
  - $E_0 = 1$
  - $t = \frac{1}{50.000} = 0{,}00002$
  - $f = 40\%$
  - $\delta(x) = x, x^2$
- Referentes a geração da consulta:
  - $c =$ falso (a diferenciação entre termos iniciais e não-iniciais feita pela normalização já cumpre o papel de dar mais importância aos termos da consulta original)
  - $\Phi(L) = \Phi_{\gamma,\lambda}(L)$, com:
    - $\lambda(x) = x, x^2, x^4, \sqrt{x}, \sqrt[4]{x}$
    - $\gamma =$ 0.5, 1, 2, 4, 8, 16, 32, 64, 128.
      Para $\lambda(x) = \sqrt{x}$ também testamos $\gamma =$ 256, 512, 1024, 2048, 4096, 8192, 16384, 32768, 65536, 131072, 232144

Grafos de Rede Contextual: $T \rightarrow T$

$MAP(\%)$		γ								
$\lambda(x)$	$\delta(x)$	0.5	1	2	4	8	16	32	64	128
x^4	x	**18.45**	17.28	16.73	16.43	16.25	16.20	16.17	16.61	16.98
x^4	x^2	**19.17**	18.50	18.02	17.38	17.16	18.05	18.46	18.91	18.84
x^2	x	**26.54**	21.77	19.80	18.49	17.47	17.18	17.32	17.62	17.56
x^2	x^2	**24.58**	21.47	19.86	18.52	18.39	19.03	19.75	20.07	19.19
x	x	34.02	34.76	34.96	35.23	**35.59**	34.77	29.86	23.67	20.15
x	x^2	35.53	35.52	35.65	**35.87**	35.84	33.04	25.27	21.80	19.93
$\sqrt{x}$	x	31.05	32.49	33.53	33.94	34.38	34.54	34.67	34.82	**35.11**
$\sqrt{x}$	x^2	33.92	34.40	34.65	34.84	34.99	35.11	35.25	35.38	**35.50**
$\sqrt[4]{x}$	x	27.53	28.74	29.76	30.86	31.68	32.28	32.87	33.32	**33.63**
$\sqrt[4]{x}$	x^2	30.56	31.64	32.24	33.02	33.70	34.03	34.23	34.17	**34.35**

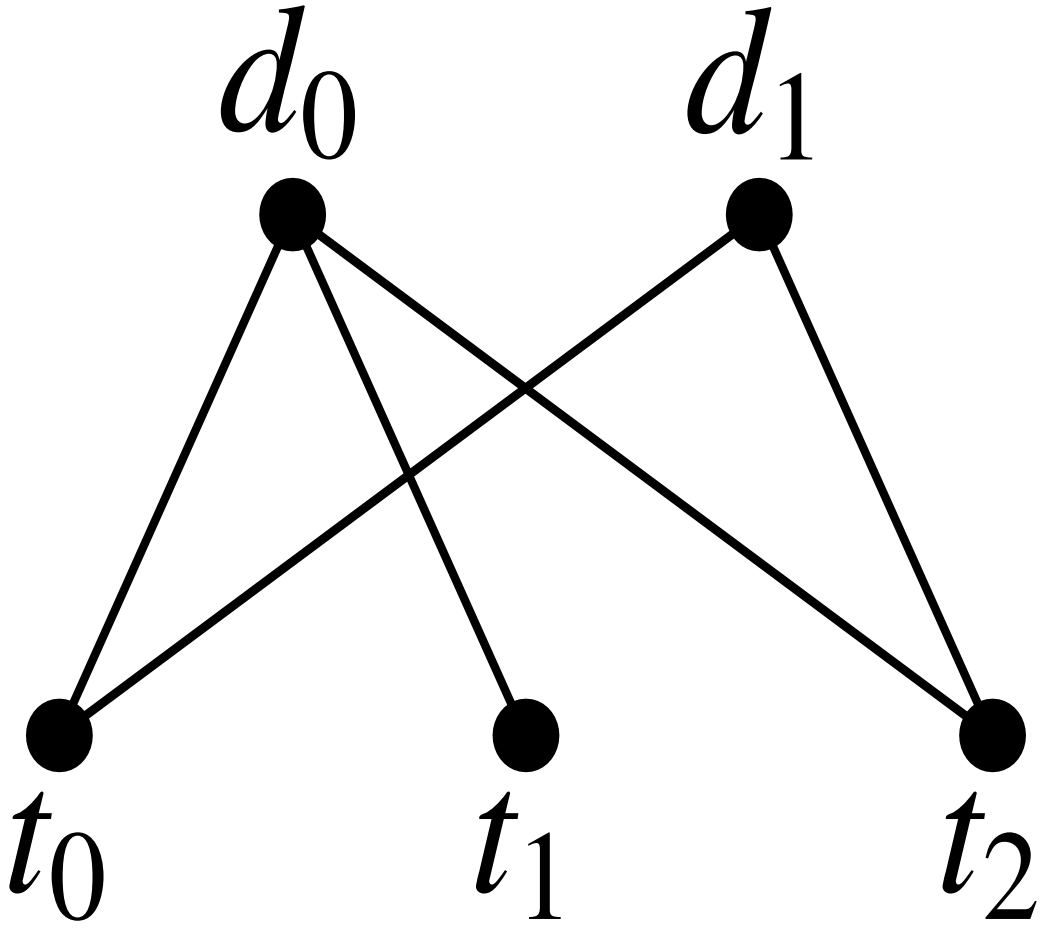
$MAP(\%)$		γ										
$\lambda(x)$	$\delta(x)$	256	512	1024	2048	4096	8192	16384	32768	65536	131072	232144
$\sqrt{x}$	x	35.29	35.45	35.61	35.75	35.88	35.98	36.06	**36.12**	36.08	35.45	33.22
$\sqrt{x}$	x^2	35.64	35.74	35.84	35.95	36.04	36.13	36.18	**36.22**	36.11	34.40	27.28
$\sqrt[4]{x}$	x											
$\sqrt[4]{x}$	x^2	34.49	34.58	34.74	34.77	34.83	34.91	34.97	35.10	35.18	35.24	**35.32**

Melhor resultado: 36,22%

Estratégia $D \rightarrow T$

- Realiza-se a pesquisa pela consulta original
- Os documentos obtidos são usados como ponto de partida da busca CNG
- Da lista de pares (termo, energia) obtidos, tomamos os Z mais energizados
- Gera-se uma nova consulta com essa lista, e realiza-se uma nova pesquisa

Estratégia $D \rightarrow T$



Estratégia $T \rightarrow T$

- Referentes a busca CNG:
  - E_0 – energia inicial
  - t – *threshold*
  - f – taxa de transmissão
  - $\delta(x)$ – função aplicada ao peso da aresta
- Referentes a geração da consulta:
  - c – booleano, define se a consulta original será adicionada
  - Z – quantos termos serão usados (usa-se os de maior energia)

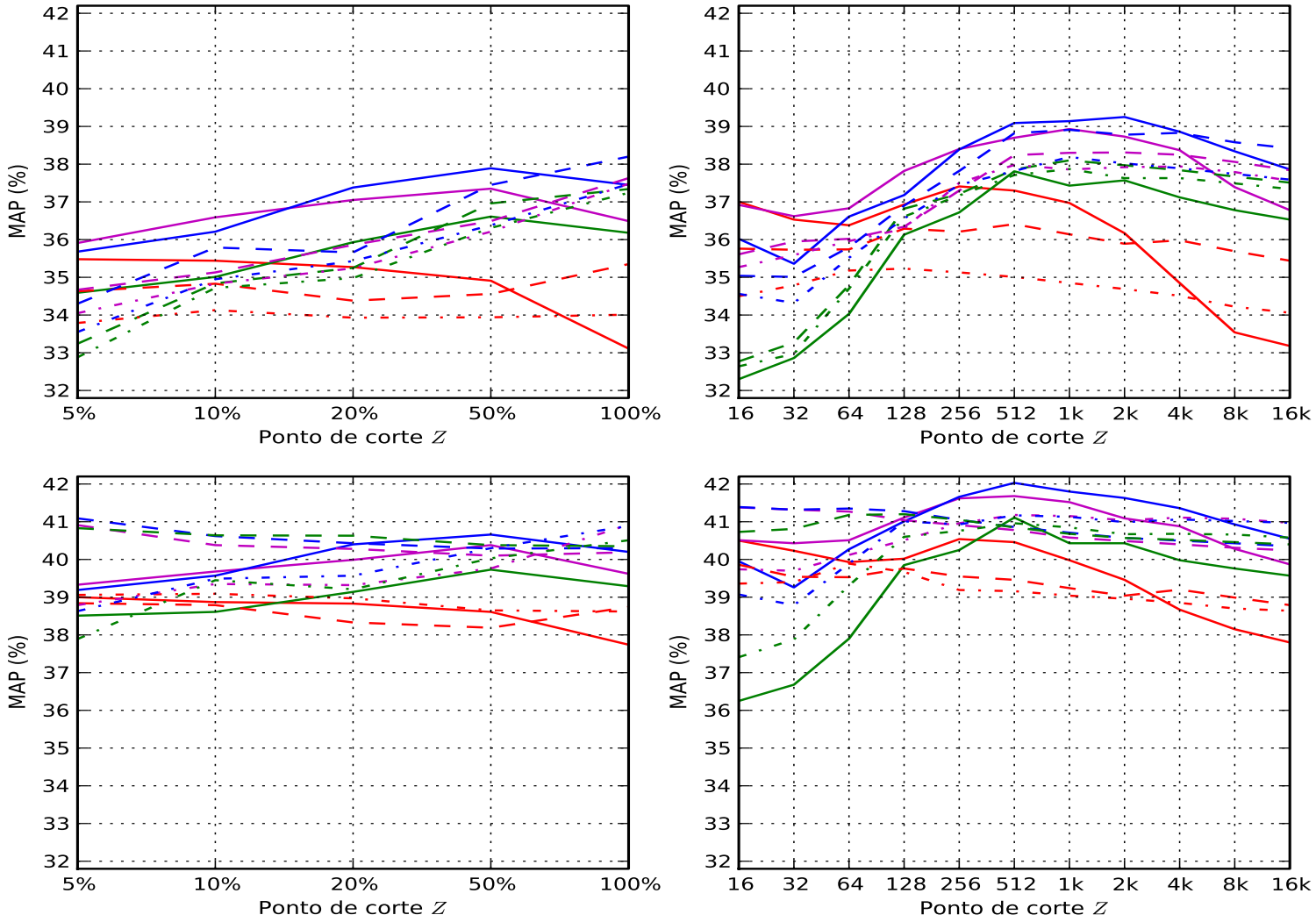
Grafos de Rede Contextual: $D \rightarrow T$

- Referentes a busca CNG:
  - $E_0 = 200$
  - $t = 1$
  - $f = 40\%, 50\%, 100\%$
  - $\delta(x) = x, x^2, x^3, x^4$
- Referentes a geração da consulta:
  - $c =$ falso, verdadeiro
  - $\Phi(L) = L$ (função identidade)
- $Z = 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 5\%, 10\%, 20\%, 50\%, 100\%$

Grafos de Rede Contextual: $D \rightarrow T$

c	$\delta(x)$	$f(\%)$	Z	MAP(%)
V	x^3	100	512	42.03
V	x^2	100	512	41.68
V	x^3	50	16	41.39
V	x^2	50	16	41.38
V	x^4	50	128	41.20
V	x^2	40	512	41.18
V	x^3	40	512	41.17
V	x^4	100	512	41.11
V	x^4	40	128	40.60
V	x	100	256	40.54
V	x	40	64	39.89
V	x	50	16	39.86
F	x^3	100	2048	39.25
F	x^2	100	1024	38.93
F	x^3	50	1024	38.91
F	x^2	50	2048	38.31
F	x^3	40	1024	38.19
F	x^4	50	1024	38.10
F	x^2	40	512	37.98
F	x^4	40	1024	37.87
F	x^4	100	512	37.81
F	x	100	256	37.41
F	x	50	512	36.41
F	x	40	128	35.23

Grafos de Rede Contextual: $D \rightarrow T$



Conclusão

CNG:

- O melhor MAP obtido pela estratégia $D \rightarrow T$ é de 42,03%
- Por sua vez, com a estratégia $T \rightarrow T$ obtivemos 36,22%
  - porém as funções de normalização usadas eram diferentes

Sintagmas Nominais:

- Técnica baseada em trabalho anterior, com algumas variações
- Antes, obteve-se 35,20% de MAP para consulta original e 29,01% para a expandida
- Agora, temos 34,01% e 44,56%, respectivamente

Conclusão e Trabalhos Futuros

- SN mostrou melhores resultados do que CNG (e tem menor custo computacional)
- Parte disso poderia ser devido à diferença de contextos (SN é local, CNG é global)
- Pretendemos em um trabalho futuro verificar se uma abordagem baseada em CNG com contexto local pode trazer melhores resultados
- Outra linha de trabalho diz respeito a SN: verificar se transformar os sintagmas em consultas por frase melhora o MAP
- Também é necessário experimentar diferentes critérios de ordenação dos sintagmas
- Finalmente, uma outra possibilidade é estudarmos formas de aliar as duas estratégias, SN e CNG, num sistema híbrido
- O melhor resultado obtido por nós (44,56%) só é inferior ao de duas das equipes participantes do CLEF, por isso acreditamos que haja espaço para melhorias