

Victor Villas Bôas Chaves

Análises de Convergência e Consistência de Métricas de Ranqueamento

Rio de Janeiro

2020

Victor Villas Bôas Chaves

Análises de Convergência e Consistência de Métricas de Ranqueamento

Dissertação submetida à Escola de Matemática Aplicada como requisito parcial para a obtenção do grau de Mestre em Modelagem Matemática da Informação.

Fundação Getúlio Vargas – FGV

Escola de Matemática Aplicada – EMap

Programa de Pós-Graduação

Rio de Janeiro

2020

Chaves, Victor Villas Bôas

Análises de convergência e consistência de métricas de ranqueamento /
Victor Villas Bôas Chaves. – 2020.
37 f.

Dissertação (mestrado) -Fundação Getulio Vargas, Escola de Matemática
Aplicada.

Orientador: Flávio Codeço Coelho.

Inclui bibliografia.

1. Sistemas de recuperação da informação. 2. Aprendizado do computador.
3. Análise matemática I. Coelho, Flávio Codeço, 1969-. II. Fundação Getulio
Vargas. Escola de Matemática Aplicada. III. Título.

CDD – 007

VICTOR VILLAS BÔAS CHAVES

"ANÁLISES DE CONVERGÊNCIA E CONSISTÊNCIA DE MÉTRICAS DE RANQUEAMENTO".

Dissertação apresentado(a) ao Curso de MESTRADO EM MODELAGEM MATEMÁTICA do(a) ESCOLA DE MATEMÁTICA APLICADA para obtenção do grau de Mestre(a) em MODELAGEM MATEMÁTICA.

Data da defesa: 21/08/2020

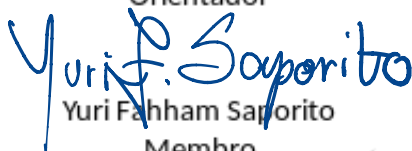
ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA

Presidente da Comissão Examinadora: Prof^o/a Flávio Codeço Coelho



Flávio Codeço Coelho

Orientador



Yuri Fahham Saporito

Membro



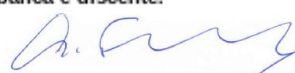
Nicolaus Linneu de Holanda

Membro

Em cumprimento ao DECRETO nº 46.970 de 13/03/20 - Poder Executivo do Estado do Rio de Janeiro, DOE nº 047A em 13/03/20, Art 4º e Portaria MEC nº 343 de 17/03/20, DOU nº 53 de 18/03/20, que dispõe sobre a suspensão temporária das atividades acadêmicas presenciais e a utilização de recursos tecnológicos (em conformidade à legislação vigente), face ao COVID-19, as apresentações das defesas de Tese e Dissertação, de forma excepcional, serão realizadas de forma remota, inclusive nessa modalidade membros da banca e discente.



César Leopoldo Camacho Manco
Diretor



Antonio de Araujo Freitas Junior
Pró-Reitor de Ensino, Pesquisa e Pós-Graduação FGV

Antonio Freitas, PhD
Pró-Reitor de Ensino, Pesquisa e Pós-Graduação
Fundação Getúlio Vargas

Instrução Normativa nº 01/19, de 09/07/19 - Pró-Reitoria FGV

Em caso de participação de Membro(s) da Banca Examinadora de forma não presencial*, o Presidente da Comissão Examinadora assinará o documento como representante legal, delegado por esta I.N.

*Skype, Videoconferência, Apps de video etc

Resumo

Cada vez mais interagimos com sistemas de buscas e esses sistemas contam com uma base crescente de documentos para filtrar e ordenar. Para avaliar a qualidade desses sistemas, diversas métricas de ranqueamento estão disponíveis com características específicas que devem encaixar no cenário de recuperação de informação. Com o advento de novas técnicas como uso de aprendizado de máquina em ranqueamento, também se espera algumas características numéricas das métricas que as tornem bons objetivos de otimização. Algumas destas métricas mais usadas como o $NDCG@k$ possuem propriedades numéricas particulares que exigem atenção, especialmente ao analisar resultados de experimentos isolados. Neste trabalho demonstramos o comportamento assintótico desta e algumas variantes de métricas de ranqueamento mais utilizadas e apresentamos uma variante do $NDCG$ que experimentalmente preserva suas características positivas enquanto mitiga características indesejadas.

Palavras-chave: recuperação da informação. ranqueamento. aprendizado de ranqueamento. métricas de ranqueamento.

Lista de ilustrações

Figura 1 – Interfaces de busca com recursos típicos	11
Figura 2 – Evolução temporal do que é considerado relevante	12
Figura 3 – DCG utilizado em sistemas de recomendação	13
Figura 4 – $P@k$ em função de N	27
Figura 5 – \bar{P} e $\bar{P}@k$ em função de N	29
Figura 6 – $NDCG$ e $NDCG@k$ em função de N	30
Figura 7 – $RDCG$ e $RDCG@k$ em função de N até 10^4	32
Figura 8 – $RDCG$ e $RDCG@k$ em função de N até 10^5	33

Lista de abreviaturas e siglas

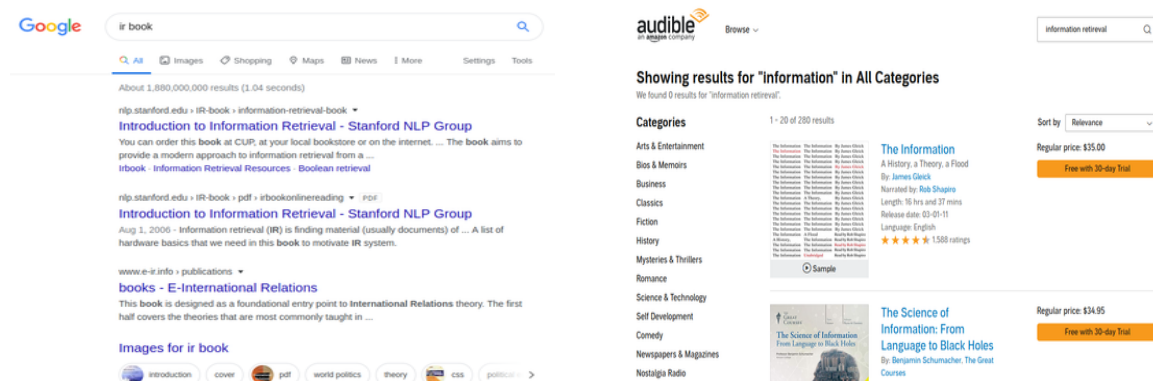
P	Precisão
$P@k$	Precisão nos top- k elementos
\bar{P}	Precisão Média
$\bar{P}@$	Precisão Média nos top- k elementos
DCG	<i>Discount Cumulated Gain</i>
$DCG@k$	<i>Discount Cumulated Gain</i> nos top- k elementos
$NDCG$	<i>Normalized Discount Cumulated Gain</i>
$NDCG@k$	<i>Normalized Discount Cumulated Gain</i> nos top- k elementos
$RDCG$	<i>Renormalized Discount Cumulated Gain</i>
$RDCG@k$	<i>Renormalized Discount Cumulated Gain</i> nos top- k elementos

Sumário

	Introdução	11
1	REFERENCIAL TEÓRICO	15
1.1	Probabilidade	15
1.2	O Problema de Ranqueamento	16
1.3	Métricas de Ranqueamento	17
1.4	A família de funções DCG	18
1.5	Distinguibilidade	19
1.6	Convergência	20
2	ANALISE SOBRE DCGS	21
2.1	Propriedades do $NDCG@k$ padrão	21
2.2	Propriedades do $NDCG$ padrão	21
2.3	DCG renormalizado	24
3	EXPERIMENTOS	25
3.1	P e $P@k$	26
3.2	\bar{P} e $\bar{P}@k$	28
3.3	$NDCG$ e $NDCG@k$	29
3.4	$RDCG$ e $RDCG@k$	31
	Considerações finais	35
	REFERÊNCIAS	37

Introdução

Sistemas de Recuperação de Informação são as ferramentas que nos permitem satisfazer uma necessidade de informação ao buscar e ranquear resultados em grande volume, atividade cada vez mais frequente e representativa das interações humanas com bases de dados (MANNING; RAGHAVAN; SCHUTZE, 2008). Enquanto a produção de conteúdo na Internet acelera e mais conteúdo antigo é catalogado, as interfaces de busca se tornam ubíquas.

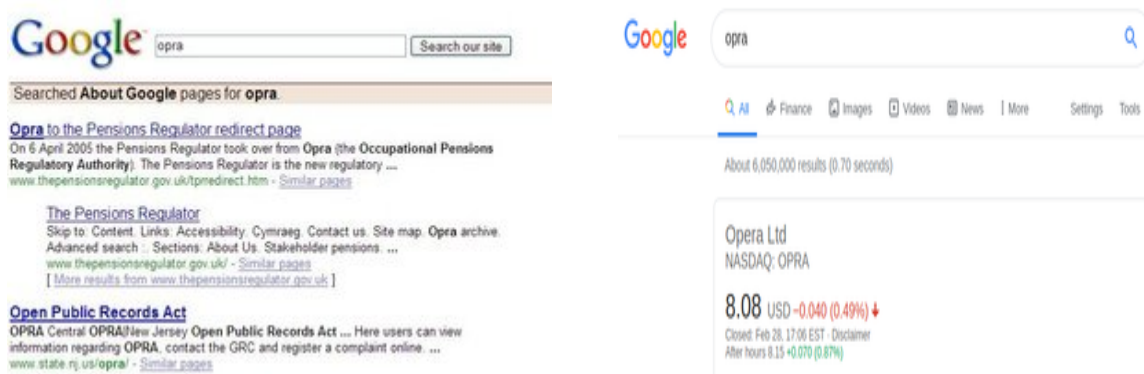


- (a) Google exibindo ocorrências dos termos pes-
quisados em negrito
- (b) Audible exibindo resultados da busca de um
subconjunto dos termos iniciais

Figura 1 – Interfaces de busca com recursos típicos

Cada sistema de busca apresenta particularidades do seu cenário de aplicação. Na figura 1a podemos ver que a sigla "IR" expressada na busca é considerada presente em uma página sobre um livro de "Relações Internacionais". Esse tipo de expansão semântica é comum em cenários de busca exploratórios que frequentemente acontecem em buscas na Web. Em contrapartida, podemos ver na figura 1b que a busca de uma loja de livros procura pelos termos exatos e ao não encontrar resultados, muda o critério de busca para poder apresentar algum resultado. Esse tipo de intervenção na intenção de busca é comum em buscas de *e-commerce* para estimular a exibição de produtos.

Apesar das muitas variações de cenários e objetivos de sistemas de recuperação de informação, um tema sempre presente é o de ranqueamento. A necessidade de priorizar resultados de interesse para o usuário trouxe a expectativa de que os resultados sempre sejam apresentados com alguma ordenação que represente a visão do sistema quanto à relevância de um documento em relação à necessidade de informação expressada (COOPER, 1968). Por exemplo, na figura 2a o sistema de busca considera relevantes apenas páginas da Web, enquanto na figura 2b o sistema considera que mostrar o preço de uma ação é prioritário e os links apresentados no sistema anterior seguem ao fim da página.



- (a) Google antigo mostrava apenas links de páginas da Web
 (b) Google novo mostra informações em tempo real no topo do ranking

Figura 2 – Evolução temporal do que é considerado relevante

Essa expectativa torna o componente de ranqueamento um tópico central na área de Recuperação de Informação (RI). Desde (ROBERTSON; BELKIN, 1978) discutia-se a importância de modelar um sistema de ranqueamento considerando a natureza complexa e multidimensional da definição de relevância, mas a própria formalização do problema de ranqueamento e a definição de graus de relevância evoluiu em conjunto das novas expectativas dos usuários.

As metodologias mais modernas posicionam o problema de ranqueamento como um problema de otimização, o que permite se fazer uso de aprendizagem de máquina e deu origem a uma nova área de pesquisa dentro de RI: *learning to rank* ou aprendizado de ranqueamento. Nesse contexto se torna essencial analisar métricas objetivo que capturem a qualidade de um sistema de recuperação de informação em função da qualidade dos rankings por ele produzidos (LIU, 2011).

Em alguns casos, métodos clássicos de estatística como classificação ou regressão foram utilizados em problemas de ranqueamento, por vezes combinados no que hoje se conhece como *regressão ordinal* (ou *classificação ordinal*). Tais métodos trazem consigo as definições usuais de métricas de erro e são amplamente estudados e seus comportamentos bem entendidos. Porém, as modelagens de maior sucesso tem sido aquelas que trazem consigo novas formas de formalizar estatisticamente os problemas de ranqueamento e reduzir a ênfase da otimização em encontrar uma aproximação de uma função, focando mais no resultado final da ordenação (LIU, 2011).

Dentre as várias métricas propostas, a família de funções *DCG* definidas em (JÄRVELIN; KEKÄLÄINEN, 2002) se destacou por seus bons resultados como função objetivo em experimentos de Aprendizado de Ranqueamento (PRADEEP; TEWARI; YANG, 2011). Sua universalidade e flexibilidade nos mais diversos cenários de RI foram bem recebidas e consolidou o padrão acadêmico e industrial para comparativos de algoritmos de ranqueamento (LIU, 2011), podendo ser encontrado em diversas especificações de

problemas de recuperação de informação e em particular, sistemas de recomendação como na figura 3.

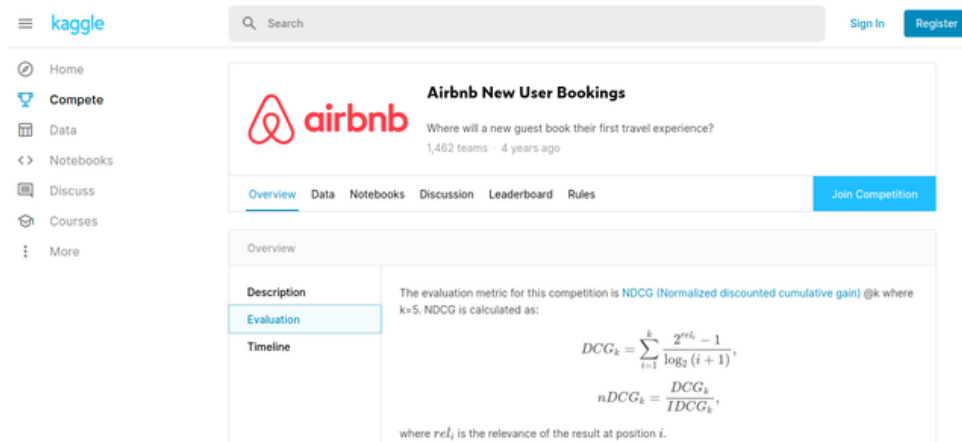


Figura 3 – DCG utilizado em sistemas de recomendação

O presente trabalho visa discutir essa família de métricas, interpretando algumas de suas propriedades e observar seus comportamentos experimentalmente. Em particular, as propriedades estudadas serão as de comportamento assintótico das métricas de ranqueamento conforme o tamanho de um ranking cresce indefinidamente. Esse comportamento não é relevante para os principais casos de uso de recuperação de informação como *web search* porque, em geral, assume-se a existência de um conjunto limitado de documentos relevantes de tamanho independente do tamanho do conjunto universo de documentos existentes. Porém, existem cenários de recuperação informação que apresentam uma característica de proporcionalidade: quanto maior o universo de documentos, maior o número de resultados desejados pelo usuário.

Como exemplo de cenário excepcional estão os fluxos de triagem gradativa de documentos: um processo de busca que envolva múltiplas etapas e participantes colaborativamente e iterativamente reduzindo o conjunto de documentos relevantes. Individualmente, um usuário pode aplicar uma busca binária em uma base de documentos de tamanho N para encontrar os $N/2$ documentos que melhor satisfazem um critério específico. De forma concreta, esse modelo pode representar um recrutador que recebeu a tarefa de selecionar metade dos candidatos para entrevistas. Investigar o comportamento assintótico das métricas em função de N pode revelar se as métricas de ranqueamento mais utilizadas cenários tradicionais de IR são adequadas para avaliar qualidades de sistemas de recuperação de informação utilizados em buscas com proporcionalidade de resultados relevantes.

1 Referencial Teórico

1.1 Probabilidade

Definição 1.1 (Colchetes de Iverson). Utilizamos a seguinte notação para indicar o resultado numérico de uma proposição lógica P :

$$[P] = \begin{cases} 1 & \text{se } P \text{ é verdadeiro} \\ 0 & \text{se } P \text{ é falso} \end{cases} \quad (1.1)$$

Definição 1.2 (Função Negligenciável). Uma função $neg : \mathbb{N} \rightarrow \mathbb{R}$ é dita negligenciável quando satisfaz a seguinte condição:

$$\forall c \in \mathbb{N}, \exists N \in \mathbb{Z} \mid |neg(x)| < \frac{1}{x^c} \quad \forall x > N \quad (1.2)$$

Definição 1.3 (Limite Superior de Conjuntos). Seja A_1, \dots, A_n, \dots uma sequência de conjuntos. O limite superior dessa sequência é o conjunto:

$$\limsup_{n \rightarrow \infty} A_n = \bigcup_{i \geq 1} \bigcap_{j \geq i} A_j \quad (1.3)$$

ou equivalentemente, é o conjunto cujos elementos são aqueles que pertencem a uma infinidade de conjuntos A_n .

Definição 1.4 (Convergência Quase Certa). Seja $(\Omega, \mathcal{F}, Pr)$ um espaço de probabilidade. Dizemos que uma sequência de variáveis aleatórias X_n converge quase certamente para X quando

$$Pr \left(\lim_{n \rightarrow \infty} X_n = X \right) = 1 \quad (1.4)$$

ou de forma equivalente:

$$Pr \left(\limsup_{n \rightarrow \infty} \{ \omega \in \Omega \mid |X_n(\omega) - X(\omega)| > \epsilon \} \right) = 0 \quad \forall \epsilon > 0 \quad (1.5)$$

Havendo a convergência, utilizamos a notação:

$$X_n \xrightarrow{q.c.} X \quad (1.6)$$

Lema 1 (Borel-Cantelli). *Seja A_1, \dots, A_n, \dots uma sequência de eventos em um espaço de probabilidade tal que*

$$\sum_{k=1}^{\infty} Pr(A_k) < \infty \quad (1.7)$$

Então é verdade também que:

$$Pr\left(\limsup_{n \rightarrow \infty} A_n\right) = 0 \quad (1.8)$$

Lema 2 (Cotas de Chernoff). *Sejam X_1, \dots, X_n variáveis aleatórias tais que $a \leq X_i \leq b$. Definindo $X = \sum_{i=1}^n X_i$ e denotando $\mu = \mathbb{E}(X)$, temos para todo $\delta > 0$:*

$$Pr(X \geq (1 + \delta)\mu) \leq e^{-\frac{2\delta^2\mu^2}{n(b-a)^2}} \quad (1.9)$$

$$Pr(X \leq (1 - \delta)\mu) \leq e^{-\frac{\delta^2\mu^2}{n(b-a)^2}} \quad (1.10)$$

ou de forma alternativa e unificando as duas equações acima:

$$Pr\left(\left|\frac{X}{\mu} - 1\right| \geq \delta\right) \leq e^{-\frac{2\delta^2\mu^2}{n(b-a)^2}} \quad (1.11)$$

1.2 O Problema de Ranqueamento

Definição 1.5 (Conjunto de Índices). O conjunto de índices para uma enumeração de tamanho n é denotado por:

$$[n] = \{k \in \mathbb{N} \mid k \leq n\} = \{1, \dots, n\} \quad (1.12)$$

Definição 1.6 (Permutação). Dado um conjunto enumerado $\mathcal{X} = \{x_1, \dots, x_n\}$, definimos uma permutação de \mathcal{X} como uma bijeção $\pi : [n] \rightarrow [n]$ que mapeia a enumeração original nos índices da nova ordenação, isto é, $\pi(i) = j$ se x_i ocupa a j -ésima posição da permutação. Podemos representar essa permutação pela lista ordenada $[x_{\pi^{-1}(1)}, \dots, x_{\pi^{-1}(n)}]$ utilizando a inversa π^{-1} e denotamos o conjunto das permutações de \mathcal{X} por $\mathfrak{S}(\mathcal{X})$.

Definição 1.7 (Ranking). Seja \mathcal{X} um conjunto de elementos $\{x_1, \dots, x_n\}$ e \mathcal{Y} o conjunto de graus de relevância $\{y_1, \dots, y_n\}$ respectivamente associados. Um *ranking* $R \in \mathfrak{S}(\mathcal{X})$ é uma permutação de \mathcal{X} cujos elementos são denotados $[r_1, \dots, r_n] = [x_{\pi^{-1}(1)}, \dots, x_{\pi^{-1}(n)}]$. Um *ranking ideal* é aquele tal que $y_{\pi^{-1}(1)} \geq \dots \geq y_{\pi^{-1}(n)}$.

A abordagem usual é assumir uma ordem crescente de prioridade conforme os graus de relevância, sendo o caso mais simples apenas $\mathcal{Y} = \{0, 1\}$ para indicar elementos irrelevantes e relevantes, respectivamente. Uma relação de ordem $y_i > y_j$ implica em um desejo de ranquear \mathcal{X} com uma permutação π de forma que $\pi(i) < \pi(j)$.

O desafio de um sistema de ranqueamento é gerar bons rankings sem conhecer *a priori* os valores verdadeiros de relevância y_i , isto é, desconhecendo a função $y(x_i) = y_i$. O que difere o problema de ranqueamento de um problema de classificação ou regressão comum em Aprendizado de Máquina é que não é necessário aproximar a função y , já que os valores y_i não são relevantes individualmente, apenas a ordem relativa (LIU, 2011).

Definição 1.8 (Função Avaliadora e Ranking Associado Canônico). Uma *função de scoring* ou função avaliadora é qualquer função $f : \mathcal{X} \rightarrow \mathbb{R}$. Um ranking $R_f = [r_1, \dots, r_n] \in \mathfrak{S}(\mathcal{X})$ diz-se associado a f se $f(r_1) \geq \dots \geq f(r_n)$. —

Não necessariamente todo sistema de ranqueamento irá explicitamente inferir uma função de avaliação e no caso mais geral pode não haver uma função de relevância explícita $y : \mathcal{X} \rightarrow [m]$. Mesmo assim, podemos usar essa formalização sem perda de generalidade pois a construção de y e f ocorre de forma implícita em qualquer sistema de ranqueamento. Essa formulação é conveniente pois a maioria dos sistemas de busca lida diretamente com funções de avaliação (WANG et al., 2013).

Em geral, é possível que f esteja associada a um conjunto de múltiplos rankings R_f associados a f , a partir de um único conjunto de elementos \mathcal{X} . Para simplificar a notação, é comum assumir a existência de um único ranking canônico que pode ser gerado deterministicamente para cada f . Por exemplo, ele pode ser obtido aplicando um algoritmo de ordenação estável sobre \mathcal{X} (orientado por f) respeitando a ordenação inicial dada pelos índices $1, \dots, n$.

1.3 Métricas de Ranqueamento

Para que seja possível comparar diferentes sistemas de ranqueamento - ou diferentes instâncias de uma função avaliadora - é necessário que eles sejam comparáveis por uma métrica que avalie a qualidade dos rankings R_f gerados. Tais métricas são denominadas na área de IR como *métricas de ranqueamento*:

Definição 1.9 (Métrica de Ranqueamento). Sejam \mathcal{X} um conjunto de elementos e \mathcal{Y} suas relevâncias associadas. Uma métrica de ranqueamento é qualquer função $\mathcal{M} : \mathfrak{S}(\mathcal{X}) \rightarrow \mathbb{R}$. Dizemos que a função f_a produz um ranking melhor que f_b segundo \mathcal{M} se e somente se

$$\mathcal{M}(R_{f_a}) > \mathcal{M}(R_{f_b}) \quad (1.13)$$

sendo R_{f_a} e R_{f_b} os respectivos rankings canônicos associados. —

Diversas métricas de ranqueamento são utilizadas na literatura e suas aplicabilidades variam conforme a natureza do problema de busca sendo priorizado. Em geral, as métricas costumam ser funções que levam em conta as posições dos elementos e podem considerar as relevâncias observadas ou não. Alguns exemplos de métricas fundamentais são:

Definição 1.10 (Precisão). Sejam \mathcal{X} um conjunto de elementos e \mathcal{Y} suas relevâncias associadas. Seja R um ranking em $\mathfrak{S}(\mathcal{X})$ e π a permutação de índices que gera R a partir de \mathcal{X} . Utilizando a notação de colchetes de Iverson (1.1), a precisão na posição k é dada por:

$$P@k(R, y_{lim}) = \frac{\sum_{i=1}^k [y_{\pi^{-1}(i)} \geq y_{lim}]}{k} \quad (1.14)$$

sendo o parâmetro y_{lim} o limiar de relevância a ser escolhido. —

Definição 1.11 (Precisão Média). Sejam \mathcal{X} um conjunto de elementos e \mathcal{Y} suas relevâncias associadas. Seja R um ranking em $\mathfrak{S}(\mathcal{X})$ e π a permutação de índices que gera R a partir de \mathcal{X} . Utilizando a notação de colchetes de Iverson (1.1), a precisão média é dada por:

$$\bar{P}@k(R, y_{lim}) = \frac{\sum_{i=1}^k P@i(R, y_{lim}) \times [y_{\pi^{-1}(i)} \geq y_{lim}]}{\sum_{i=1}^k [y_{\pi^{-1}(i)} \geq y_{lim}]} \quad (1.15)$$

efetivamente calculando a média dos valores de precisão $P@k$ para k em que os elementos relevantes aparecem. —

É comum que métricas de ranqueamento sejam parametrizadas por um índice k que representa a posição máxima de interesse para a métrica. A natureza das aplicações de ranqueamento costuma resultar em conjuntos de elementos \mathcal{X} muito grandes e pouca importância quanto à qualidade do ranking em posições longe do início (LIU, 2011). Ao omitir k , deve-se assumir $k = |\mathcal{X}|$.

Com o avanço das técnicas de Aprendizagem de Ranqueamento, mais ênfase tem se dado em observar o comportamento dessas métricas que costumam ser descontínuas e não diferenciáveis em função das relevâncias - características importantes para a otimização convexa baseada em gradientes que atualmente predominam em LTR (LIU, 2011).

1.4 A família de funções DCG

Definição 1.12 (DCG e $NDCG$). Sejam \mathcal{X} um conjunto de elementos e \mathcal{Y} suas relevâncias associadas. Seja R um ranking em $\mathfrak{S}(\mathcal{X})$, π a permutação de índices que gera R a partir de \mathcal{X} e π^{-1} a permutação inversa de π . A métrica de erro DCG (*Discounted Cumulated Gain* ou Ganho Acumulado Descontado) e sua métrica complementar $NDCG$ (*Normalized DCG*) são definidas pela equações:

$$DCG@k(R) = \sum_{i=1}^k G(y_{\pi^{-1}(i)}) D(i) \quad (1.16)$$

$$NDCG@k(R) = \frac{DCG@k(R)}{\max_{S \in \mathfrak{S}(\mathcal{X})} DCG@k(S)} \quad (1.17)$$

sendo os hiper-parâmetros funcionais G para o Ganho e D para o Desconto em função do índice i de cada elemento e suas respectivas relevâncias y_i . O $NDCG$ é obtido normalizando o DCG em relação ao valor máximo ideal para esses elementos. —

A motivação por trás da formulação do DCG explicada pelos autores originais em (JÄRVELIN; KEKÄLÄINEN, 2002) é de modelar o ganho acumulado que um usuário vivencia ao consumir os elementos da listagem até a posição k . Essa métrica se destaca das principais soluções anteriores (como a precisão e a precisão média) por implementar a sugestão apresentada em (ROBERTSON; BELKIN, 1978) de colocar a ênfase das métricas de ranqueamento na experiência de consumo da listagem do ponto de vista do usuário.

Dentro dessa família de funções, o uso de ganho $G(y) = y$ linear e desconto $D(i) = 1/\log(1+i)$ logarítmico se estabeleceu como padrão da área de *Web Search* (LIU, 2011), ocasionalmente sendo usado ganho $G(y) = 2^y$ exponencial como em (BURGES et al., 2005). Essa configuração de ganho linear e desconto logarítmico é referida por (LIU, 2011) como *NDCG padrão* e deve-se assumir o caso quando não for especificado o contrário.

A formulação frequentemente utilizada em *Web Search* é limitar o somatório do DCG com ênfase em valores de k entre 1 e 10. Essa ênfase se deve ao fato de que nessas aplicações de RI a qualidade do ranqueamento para elementos longe do topo é irrelevante, sendo predominante nos principais trabalhos na área de Aprendizado de Ranqueamento como em (WANG et al., 2017). Seguindo a conveção das métricas anteriores, quando k for omitido deve-se supor $k = |\mathcal{X}|$.

1.5 Distinguibibilidade

Uma caracterização para métricas de ranqueamento apresentada por (WANG et al., 2013) se refere ao comportamento assintótico de uma métrica de ranqueamento ao comparar duas funções avaliadoras.

Definição 1.13 (Distinguibibilidade Consistente). Seja $S_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ um conjunto de pares de elementos e relevâncias *i.i.d.* amostradas em $X \times Y$ com distribuição P_{XY} . Sejam f_a e f_b duas funções avaliadoras com rankings associados canônicos R_{f_a} e R_{f_b} . Dizemos que f_a e f_b são *consistentemente distinguíveis* por uma métrica de ranqueamento \mathcal{M} quando existe uma função negligenciável $neg(N)$ (1.2) tal que para todo N suficientemente grande, com probabilidade $1 - neg(N)$ temos

$$\mathcal{M}(R_{f_a}) \geq \mathcal{M}(R_{f_b}) \quad \forall n \geq N \quad (1.18)$$

ou

$$\mathcal{M}(R_{f_a}) \leq \mathcal{M}(R_{f_b}) \quad \forall n \geq N \quad (1.19)$$

Essa característica é atraente visto que em alguns domínios é comum encontrar rankings com milhões ou bilhões de elementos para ordenar, como é o caso com Web Search. Apesar de ser importante que a métrica de ranqueamento priorize o desempenho do sistema nos primeiros resultados, é conveniente que sejamos capazes de comparar esses sistemas quando o comportamento assintótico conforme o número de elementos cresce. Isto é, gostaríamos que a métrica de ranqueamento não fosse sensível ao tamanho do conjunto a ranquear de forma a atrapalhar a comparação entre duas funções avaliadoras para conjuntos arbitrariamente grandes.

1.6 Convergência

Métricas complexas podem ser difíceis de provar consistentemente distinguíveis, mas uma outra forma de tentar observar o comportamento assintótico é investigar sua convergência quase certa. A intuição é de que se $M(R_{f_a})$ e $M(R_{f_b})$ convergem para pontos distintos, então elas serão distinguíveis. Se a convergência ocorre no mesmo ponto, é possível ter distinguibilidade consistente ou não.

Se a convergência é um indicativo de distinguibilidade consistente, pode parecer a princípio que a distinguibilidade é um critério excessivamente exigente. Porém, se pensarmos que estamos analisando rankings arbitrariamente longos, temos como entrada amostras S_n de P_{XY} arbitrariamente volumosas e portanto, podemos estimar com precisão arbitrária a distribuição P_{XY} . Portanto, intuitivamente deveríamos ser capazes de reduzir a variância da métrica observada que possa depender das relevâncias disponíveis na amostragem e nos aproximar de uma medida pura da qualidade da função avaliadora f . O critério de convergência significa apenas que a métrica deveria ser capaz de aproveitar o conhecimento completo de P_{XY} e portanto ser capaz de consistentemente medir a qualidade das funções avaliadoras sem interferência da amostragem.

2 Analise sobre DCGs

2.1 Propriedades do $NDCG@k$ padrão

Ao tentar caracterizar quais funções de desconto D seriam possíveis de se usar no $NDCG$ a fim de obter a propriedade de distinguibilidade consistente, (WANG et al., 2013) mostrou pela primeira vez que qualquer função de desconto que decresça ao ponto da convergência em somatório faz com que $NDCG(f, S_n)$ não seja consistentemente distinguível. Isto é:

Teorema 1 (Divergência e Inconsistência para $NDCG@k$). *Para $G(y) = y$, se $\sum_{i=1}^{\infty} D(i) < \infty$, então o $NDCG$ não converge e não é consistentemente distinguível.* —

Intuitivamente, se a função de desconto decresce rápido a ponto de a cauda do ranking pouco contribuir para a métrica DCG , então um conjunto \mathcal{X} suficientemente grande introduz uma variabilidade de relevâncias disponíveis tão grande que a métrica oscila muito e a probabilidade de $NDCG@k(R_{f_a}) \geq NDCG@k(R_{f_b})$ permanece significante mesmo que na média tenhamos f_a melhor que f_b .

O objetivo inicial do estudo das funções admissíveis D para obter distinguibilidade consistente é justificar o uso de desconto logarítmico $D(i) = 1/\log(1+i)$ ao invés de um desconto quadrático como $D(i) = 1/i^2$, que pelo Teorema 1 gera uma métrica inconsistente. Porém, é fácil observar que o $NDCG@k$ padrão também encaixa no Teorema 1 pois para k fixo, temos:

$$\sum_{i=1}^{\infty} D(i) = \sum_{i=1}^k D(i) < \infty \quad (2.1)$$

A única forma de fazer $NDCG@k$ consistentemente distinguível é utilizar para k uma função crescente em n , tal que o somatório da equação (2.1) divirja. De fato, (WANG et al., 2013) mostra que para $k = cn$ com $c \in (0, 1]$, o $NDCG@k$ padrão é uma métrica consistentemente distinguível.

2.2 Propriedades do $NDCG$ padrão

Antes de discutir a distinguibilidade consistente do $NDCG$ padrão, é interessante observar sua convergência. O Teorema 1 implica que o $NDCG@k$ não converge, mas não somente temos convergência para $NDCG$ como ela ocorre para seu valor máximo:

Teorema 2 (Convergência do $NDCG$). *Assumindo uma amostra de documentos e relevâncias $S_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ amostrados i.i.d. com distribuição P_{XY} sobre $X \times Y$, com então para toda função de avaliação $f : \mathcal{X} \rightarrow \mathbb{R}$ temos para $n \rightarrow \infty$:*

$$NDCG(R_f) \xrightarrow{q.c.} 1 \quad (2.2)$$

Demonstração. Seja $l = \sum_{i=1}^n y_i$ o total agregado de relevância em um ranking, com $m = \max(Y)$ e $k = \lfloor \frac{l}{m} \rfloor$. Para l fixo, o melhor caso de permutação $\bar{\pi}$ é gerar um ranking com k documentos de relevância $z_{\bar{\pi}} = m$ seguidos de 1 documento de relevância $z_{\bar{\pi}} = km - l$ seguido de $n - k - 1$ documentos de relevância $z_{\bar{\pi}} = 0$, enquanto o pior caso de permutação π é gerar um ranking exatamente o reverso do gerado por $\bar{\pi}$. Nessas condições, temos:

$$NDCG(R_f) \geq \frac{\sum_{i=1}^n z_{\pi^{-1}(i)} / \log(1+i)}{\sum_{i=1}^n z_{\bar{\pi}^{-1}(i)} / \log(1+i)} \quad (2.3)$$

$$\geq \frac{\sum_{i=n-k+1}^n m / \log(1+i)}{\sum_{i=1}^{k+1} m / \log(1+i)} \quad (2.4)$$

$$= \frac{\sum_{i=1}^n 1 / \log(1+i) - \sum_{i=1}^{n-k} 1 / \log(1+i)}{\sum_{i=1}^{k+1} 1 / \log(1+i)} \quad (2.5)$$

Utilizando a função logaritmo integral ajustada $Li(n) = \int_2^n \frac{dx}{\log(x)}$ e reconhecendo que $Li(n)$ é uma função côncava ($Li''(n) \leq 0$), podemos usar a soma de Riemann pelo ponto médio como um limite inferior e a soma de Riemann à direita como limite superior

$$\sum_{i=1}^n \frac{1}{\log(1+i)} - \underbrace{\frac{\frac{1}{\log(1+n)} + \frac{1}{\log(1+1)}}{2}}_{o(1)} \leq \int_1^n \frac{dx}{\log(1+x)} = Li(n+1) \leq \sum_{i=2}^n \frac{1}{\log(1+i)}$$

e de (2.5) chegamos em:

$$NDCG(R_f) \geq \frac{Li(n+1) - (Li(n-k+1) + o(1))}{Li(k+2) + o(1)} \quad (2.6)$$

$$= \frac{Li(n+1) - Li(n-k+1)}{Li(k+2)} - o(1) \quad (2.7)$$

onde o símbolo de Landau o (o-pequeno) representa uma função cujo comportamento assintótico é dominado por outra. Em particular, f é $o(1)$ se e somente se $\lim_{x \rightarrow \infty} f(x) = 0$.

Para prosseguir, podemos aplicar as Cotas de Chernoff às variáveis aleatórias $0 \leq y_i \leq m$, com $l = \sum_{i=1}^n y_i$ e $\mu = \mathbb{E}[l]$, obtendo:

$$Pr \left(\left| \frac{l}{\mu} - 1 \right| \geq \delta \right) \leq 2e^{-2 \frac{\delta^2 \mu^2}{nm^2}} \quad (2.8)$$

Se condicionarmos no evento de $\left|\frac{l}{\mu} - 1\right| < \delta$, denotando $\nu = \mathbb{E}[y_i]$, temos $\mu = n\nu$ e podemos dizer que:

$$(\nu - \delta\nu)n \leq l \leq (\nu + \delta\nu)n \quad (2.9)$$

Da definição de k temos $km \leq l \leq (k+1)m - 1$ e de (2.9):

$$((\nu - \delta\nu)n + 1)m^{-1} - 1 \leq k \leq (\nu + \delta\nu)nm^{-1} \quad (2.10)$$

Substituindo (2.10) em (2.7), temos:

$$NDCG(R_f) \geq \frac{Li(n+1) - Li(n - ((\nu - \delta\nu)n + 1)m^{-1} + 2)}{Li((\nu + \delta\nu)nm^{-1} + 2)} - o(1) \quad (2.11)$$

$$\geq \frac{Li(n+1) - Li((1 - (\nu - \delta\nu)m^{-1})n + 2 - m^{-1})}{Li((\nu + \delta\nu)m^{-1}n + 2)} - o(1) \quad (2.12)$$

Sabendo que $Li(n+c) \sim Li(n) \sim \frac{n}{\log(n)}$ para $n \rightarrow \infty$ e supondo que podemos escolher δ tal que $\delta \rightarrow 0$, temos condicionado ao evento $\left|\frac{l}{\mu} - 1\right| < \delta$:

$$NDCG(R_f) \geq \frac{\frac{n}{\log(n)} - \frac{(1 - (\nu - \delta\nu)m^{-1})n}{\log((1 - (\nu - \delta\nu)m^{-1})n)}}{\frac{(\nu + \delta\nu)m^{-1}n}{\log((\nu + \delta\nu)m^{-1}n)}} - o(1) \quad (2.13)$$

$$= \frac{\frac{1}{\log(n)} - \frac{1 - (\nu - \delta\nu)m^{-1}}{\log((1 - (\nu - \delta\nu)m^{-1})n)}}{\frac{(\nu + \delta\nu)m^{-1}}{\log((\nu + \delta\nu)m^{-1}n)}} - o(1) \quad (2.14)$$

$$= \frac{\frac{1}{\log(n)} - \frac{1 - \nu m^{-1}}{\log(n)}}{\frac{\nu m^{-1}}{\log(n)}} - o(1) \quad (2.15)$$

$$= \frac{1}{\nu m^{-1}} - \frac{1 - \nu m^{-1}}{\nu m^{-1}} - o(1) \quad (2.16)$$

$$= 1 - o(1) \quad (2.17)$$

Portanto, para todo $\epsilon > 0$ existe N suficientemente grande tal que para $n > N$ temos:

$$Pr(|NDCG(R_f) - 1| > \epsilon) \leq 2e^{-2\frac{\delta^2 \nu^2 n}{m^2}} \quad (2.18)$$

Escolhendo $\delta = \frac{n^{-1/3}m}{\nu}$ é suficiente para satisfazer $\delta \rightarrow 0$ quando $n \rightarrow \infty$ e obter:

$$Pr(|NDCG(R_f) - 1| > \epsilon) \leq 2e^{-2n^{1/3}} \quad (2.19)$$

Observando que $\sum_{n=1}^{\infty} 2e^{-2n^{1/3}} < \infty$, pelo lema de Borel-Cantelli temos:

$$Pr(\limsup_{n \rightarrow \infty} |NDCG(R_f) - 1| > \epsilon) = 0 \quad (2.20)$$

provando a convergência quase certa. \square

A demonstração generaliza a apresentada em (WANG et al., 2013) feita para $m = 1$. Nota-se que o procedimento de prova independe das propriedades de f , emergindo apenas do comportamento assintótico da função $NDCG$ e da particularidades do melhor e pior caso de ordenação.

A princípio, poderíamos interpretar o Teorema 2 como evidência de que rankings suficientemente grandes tendem a tornar as funções avaliadores indistinguíveis, dado que em todo caso o $NDCG$ tende convergir para o valor máximo. Apesar da convergência, (WANG et al., 2013) demonstrou que o $NDCG$ é consistentemente distinguível. Portanto, apesar de o Teorema 2 dizer que para rankings arbitrariamente grandes qualquer função avaliadora irá convergir para o $NDCG$ máximo, (WANG et al., 2013) mostra que as melhores funções avaliadoras são as que fazem o $NDCG$ convergir mais rapidamente.

2.3 DCG renormalizado

Uma nova proposta de normalização do DCG presente em (KATERENCHUK; ROSENBERG, 2018) normaliza o DCG não apenas considerando seu valor máximo, mas também seu valor mínimo. O objetivo dessa normalização é que seja possível argumentar que uma ordenação é boa ou ruim observando apenas se a métrica está próxima de 1 ou 0, respectivamente - o que não é possível no $NDCG$.

Para não confundir com a nomenclatura original de $NDCG$ como o DCG normalizado e como não iremos utilizar a métrica apresentada em (KATERENCHUK; ROSENBERG, 2018) (que apresenta mais alterações do que o modo de normalização), vamos chamar essa nova métrica derivada de DCG renormalizado ou $RDCG$.

Definição 2.1 ($RDCG$).

$$RDCG(R) = \frac{DCG(R) - \min_{L \in \mathfrak{G}(\mathcal{X})} DCG(L)}{\max_{L \in \mathfrak{G}(\mathcal{X})} DCG(L) - \min_{L \in \mathfrak{G}(\mathcal{X})} DCG(L)} \quad (2.21)$$

Como temos $RDCG(R) = 0$ no pior caso de ordenação reversa da ordenação ideal, não é possível utilizar o mesmo caminho de prova feito para o Teorema 2. Provar uma possível convergência do $RDCG$ permanece em aberto e sendo uma possível direção de trabalhos futuros.

3 Experimentos

Diversos experimentos foram feitos para observar o comportamento assintótico das métricas de ranqueamento analisadas. Para gerar os graus de relevância, foi usada como base uma distribuição Gaussiana ($\mu = 0$ e $\sigma = 3$), geramos amostragens de diversos tamanhos (de 10^2 a 10^4 , dependendo da velocidade de convergência da métrica analisada). Em seguida, arredondamos cada valor para o inteiro mais próximo e saturamos no intervalo $[0, 10]$ forçando os valores < 0 para 0 e > 10 para 10. Não aplicamos nenhuma ordenação, o que pode ser representado por uma função avaliadora f constante. Em seguida, colocamos num gráfico os valores da métrica de ranqueamento em função do tamanho do ranking. Para analisar a convergência, cada N gera 100 rankings diferentes e no gráfico representamos a média e o intervalo dado pelo desvio padrão ($\pm 1\sigma$) em cada N .

Os parâmetros escolhidos geram listas de relevâncias com uma concentração acentuada de documentos pouco relevantes, porém com uma quantidade considerável de documentos muito relevantes. Ao manter a distribuição de probabilidade constante para diferentes tamanhos de ranking, simulamos uma característica de proporcionalidade: a quantidade esperada de documentos com cada grau relevância se mantém. Ao escolher uma função avaliadora f constante, também concretizamos experimentalmente os cenários em que a métrica contra-intuitivamente converge para seu valor máximo.

Como exemplo, o código que gera as análises de convergência para o *NDCG* está exposto na listagem 3.1. Os programas que geram as outras análises são semelhantes e para manter a brevidade do trabalho, estão acessíveis publicamente em <https://github.com/villasv/ir-experiments>.

Código 3.1 – Experimento de convergência do NDCG

```
import random

import numpy as np
import pandas as pd
import seaborn as sns

... metric defitions ...

random.seed(42)
ndcg_nn = NDCG()
ndcg_05 = NDCG(k=5)
ndcg_50 = NDCG(k=50)
```

```

def gen_ranking(n):
    return np.clip(np rint(np.random.normal(0, 3, n)), 0, 10)

def main():
    sampling = 100
    linspace = np.linspace(10**2, 10**4, num=100)
    ranksize = [int(i) for i in np rint(linspace)]
    rankings = [
        gen_ranking(n)
        for n in ranksize
        for _ in range(sampling)
    ]

    df = pd.DataFrame(
        [[r.size, 'N', ndcg_nn(r)] for r in rankings] +
        [[r.size, '05', ndcg_05(r)] for r in rankings] +
        [[r.size, '50', ndcg_50(r)] for r in rankings],
        columns=['N', 'k', 'NDCG@k']
    )
    sns.lineplot(
        x='N', y='NDCG@k', hue='k', ci='sd', data=df,
    ).get_figure().savefig('experiment.png')

if __name__ == "__main__":
    main()

```

3.1 P e $P@k$

Neste experimento foi utilizada a implementação de $P@k$ exibida no código 3.3.

Código 3.2 – Implementação de $P@k$

```

class PatK(Metric):
    def __init__(self, ylim, k=None):
        self._ylim = ylim

```

```

self._k = k

def _eval(self, r):
    r = np.asarray(r)
    k = self._k if self._k is not None else r.size
    rk = r[:k]
    return np.sum(rk >= self._ylim) / k

```

Podemos observar na Figura 4 que a Precisão $P@k$ definida em (1.10) não converge nem reduz em variabilidade conforme N cresce. Essa é uma forte evidência contra seu poder de distinguibilidade e sua aplicabilidade em análises estatísticas.

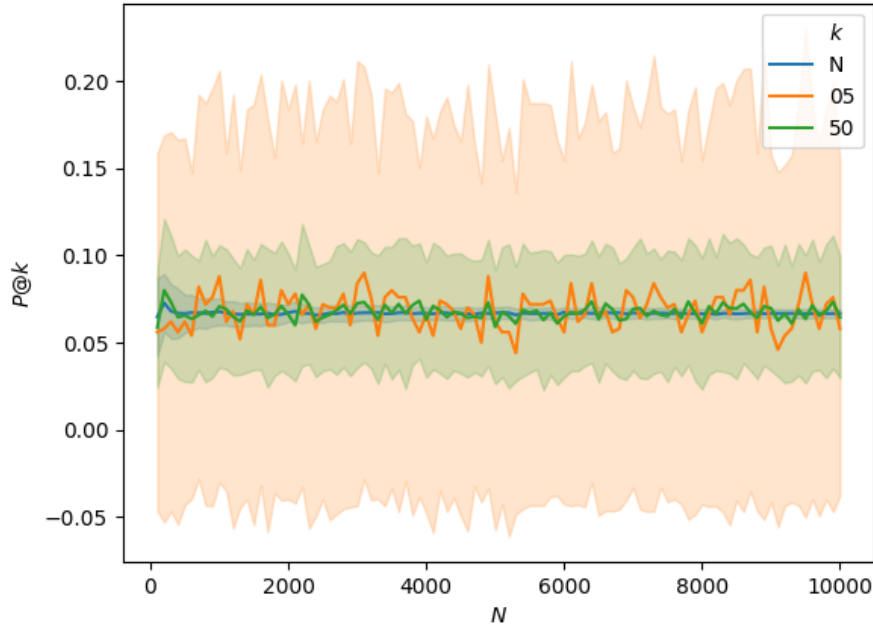


Figura 4 – $P@k$ em função de N

Também podemos observar que maiores valores de k reduzem o desvio padrão, conforme esperávamos pela intuição de que quanto mais ênfase nos poucos primeiros resultados, maior variabilidade da métrica.

Vemos que fazendo $k = N$ a precisão parece convergir, mas a métrica em si perde a utilidade. Ela apenas mede quantos documentos são relevantes (independente de ordem) e é mais uma métrica da qualidade da distribuição P_{XY} do que da função avaliadora f e, apesar de convergir, não é uma boa métrica de ranqueamento.

Por fim, podemos observar que para $k = N$ ou para k constante, neste experimento a média das métricas tende a coincidir. Isso se deve ao fato de que a ordenação está completamente aleatória e qualquer precisão medida (independente de k) está estimando

a qualidade da distribuição P_{XY} .

3.2 \bar{P} e $\bar{P}@k$

Neste experimento foi utilizada a implementação de $\bar{P}@k$ exibida no código 3.3.

Código 3.3 – Implementação de $\bar{P}@k$

```
class AveP(Metric):
    def __init__(self, ylim, k=None):
        self.__ylim = ylim
        self.__k = k

    def __eval(self, r):
        r = np.asarray(r)
        k = self.__k if self.__k is not None else r.size
        rk = r[:k]
        n_pos = np.sum(rk >= self.__ylim)
        if n_pos < 1:
            return 0

        patks = 0
        for i in range(k):
            if rk[i] >= self.__ylim:
                patks += PatK(self.__ylim, i+1)(rk)
        return patks / n_pos
```

Imediatamente podemos observar na Figura 5 que diferentemente do que acontece para $P@k$, \bar{P} converge para um valor diferente de onde $\bar{P}@k$ oscila. Isso indica que de fato algo em relação à ordenação está sendo medido, já que o valor da métrica se desloca para fora da região de oscilação para amostras fixas. Ao contrário da precisão, onde não faz sentido usar $k = N$, a precisão média leva em conta as posições dos itens relevantes e pode ser utilizada no ranking inteiro.

Podemos concluir que \bar{P} possui melhores propriedades do que P , mas $\bar{P}@k$ possui os mesmos problemas de convergência esperados para métricas que descartam a cauda do ranking. Também podemos concluir que \bar{P} é uma métrica mais útil do que P ($P@k$ com $k = N$) pois não mede simplesmente a qualidade da amostragem, mas leva em conta a ordenação. Porém, vemos que o valor de convergência é extremamente baixo (< 0.1) assim como para P , pois a métrica aplica penalidades fortes para itens relevantes longe do topo

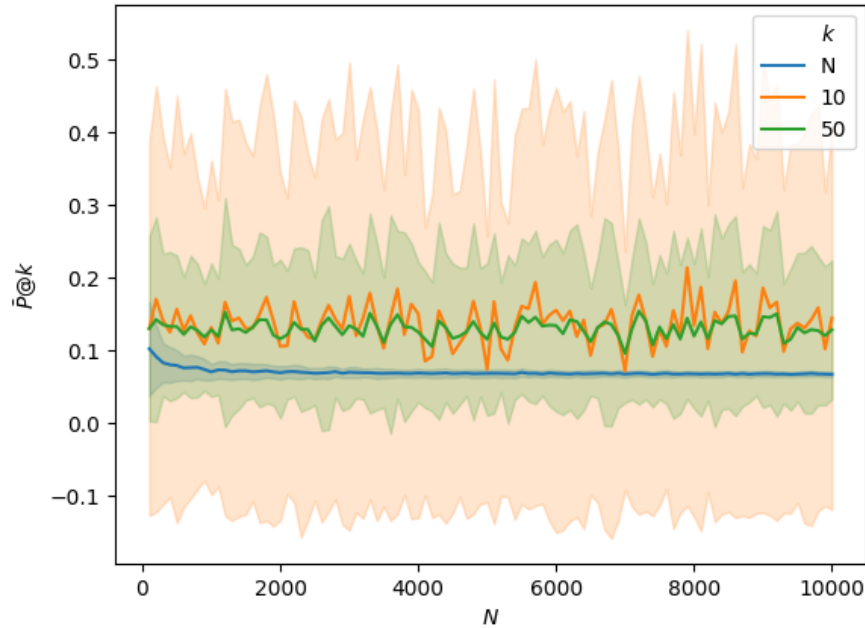


Figura 5 – \bar{P} e $\bar{P}@k$ em função de N

do ranking, algo que é esperado de ocorrer em rankings muito grandes, portanto a métrica é pouco aplicável em sistemas de busca com grandes volumes.

3.3 *NDCG* e *NDCG@k*

Neste experimento foi utilizada a implementação de *NDCG@k* exibida no código 3.4.

Código 3.4 – Implementação de *NDCG@k*

```
class DCG(Metric):
    def __init__(self, k=None, G=identity, D=p1log2):
        self._k = k
        self._G = G
        self._D = D

    def _eval(self, r):
        r = np.asarray(r)
        k = self._k if self._k is not None else r.size
        rk = r[:k]
        return np.sum(self._G(rk) / self._D(np.arange(1, k+1)))
```

```

class NDCG(Metric):
    def __init__(self, k=None, G=identity, D=p1log2):
        self._dcg = DCG(k, G, D)

    def _eval(self, r):
        max_dcg = self._dcg(sorted(r, reverse=True))
        return max_dcg and (self._dcg(r) / max_dcg)

```

Podemos observar na Figura 6 que o $NDCG@k$ definido em (1.10) não aparenta reduzir seu desvio padrão e convergir, confirmando o Teorema 1. Aumentar o valor de k parece reduzir o desvio padrão, também corroborando com o que sabemos sobre k poder ser menor que N , desde que cresça de forma ilimitada para $NDCG@k$ convergir.

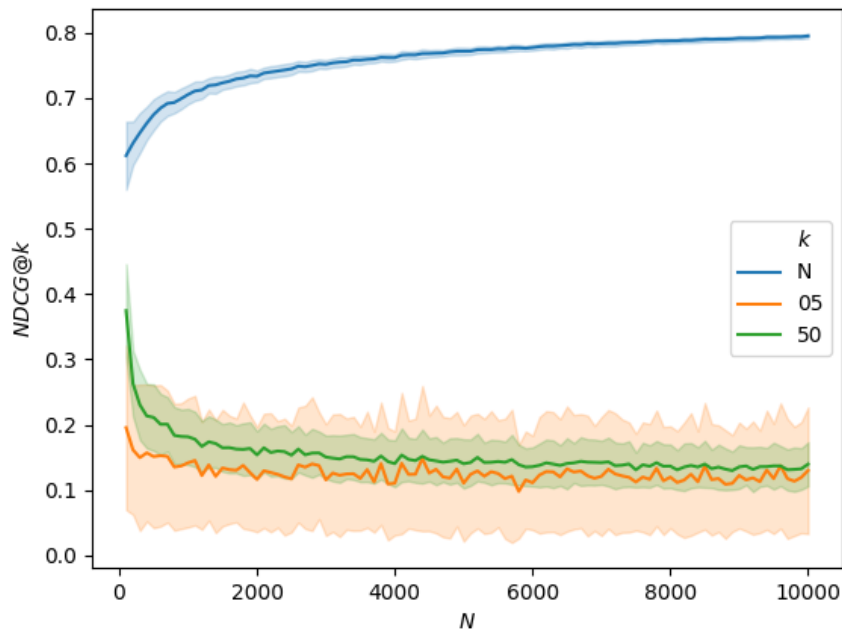


Figura 6 – $NDCG$ e $NDCG@k$ em função de N

Ao mesmo tempo, para o $NDCG$ o desvio padrão aparenta decrescer rapidamente e a métrica convergir em direção a 1, conforme o esperado pelo Teorema 2. Mais que isso, o desvio padrão já é extremamente pequeno para $N = 5000$ e já para $k = 10000$ uma função avaliadora trivial consegue chegar próximo de $NDCG = 0.8$.

Essa rápida convergência do $NDCG$ para próximo de 0.8 para rankings ainda longe de chegar na ordem de um milhão de documentos mostra que é necessário ter cuidados ao utilizar essa métrica:

- Em experimentos menos controlados (e.g. acompanhamento de um sistema já implan-

tado), para que possamos ter certeza de que uma variação do *NDCG* se deve a uma melhoria da função de avaliação f , precisamos primeiro garantir que os tamanhos dos rankings observados são os mesmos;

- Os valores do *NDCG* observados isoladamente não são interpretáveis. Alcançar um $NDCG = 0.8$ ranqueando 50 documentos exige uma boa função avaliadora, enquanto alcançar o mesmo valor com milhões de documentos é trivial. Análises de qualidade baseadas no *NDCG* devem ser feitas comparativamente a uma segunda função avaliadora f (aproveitando sua propriedade de distinguibilidade consistente).

3.4 *RDCG* e *RDCG@k*

Neste experimento foi utilizada a implementação de *RDCG@k* exibida no código 3.5.

Código 3.5 – Implementação de *RDCG@k*

```
class RDCG(Metric):
    def __init__(self, k=None, G=identity, D=p1log2):
        self._dcg = DCG(k, G, D)

    def _eval(self, r):
        max_dcg = self._dcg(sorted(r, reverse=True))
        min_dcg = self._dcg(sorted(r))
        a = (self._dcg(r) - min_dcg)
        b = (max_dcg - min_dcg)
        g['DONES'] += 1
        print(f"{g['DONES']}/{g['TOTAL']}")
        return a / b if b > 0 else 0
```

Na Figura 7 podemos ver que assim como nas métricas anteriores, a versão limitada até os top- k itens reduz o desvio padrão em função de k , mas com k fixo não converge. Assim como o *NDCG*, o *RDCG* também aparenta convergir.

A convergência do *RDCG* aparenta ser significativamente mais lenta para essa mesma função avaliadora f constante. Diferentemente dos outros experimentos, nesse caso decidimos fazer um experimento extra com rankings de até 10^5 elementos. Com 10^4 a redução do desvio padrão não é tão aparente, mas na Figura 8 é mais visível.

Apesar de não ter sido possível demonstrar o poder de distinguibilidade consistente do *RDCG*, os resultados dos experimentos de convergência indicam que a métrica possui boas propriedades numéricas. Em particular, os valores absolutos da métrica são mais

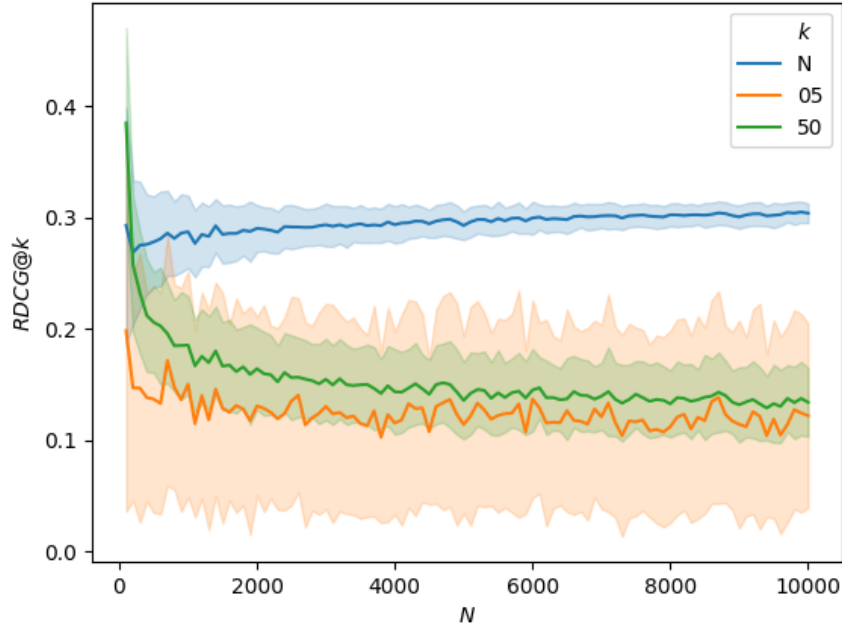
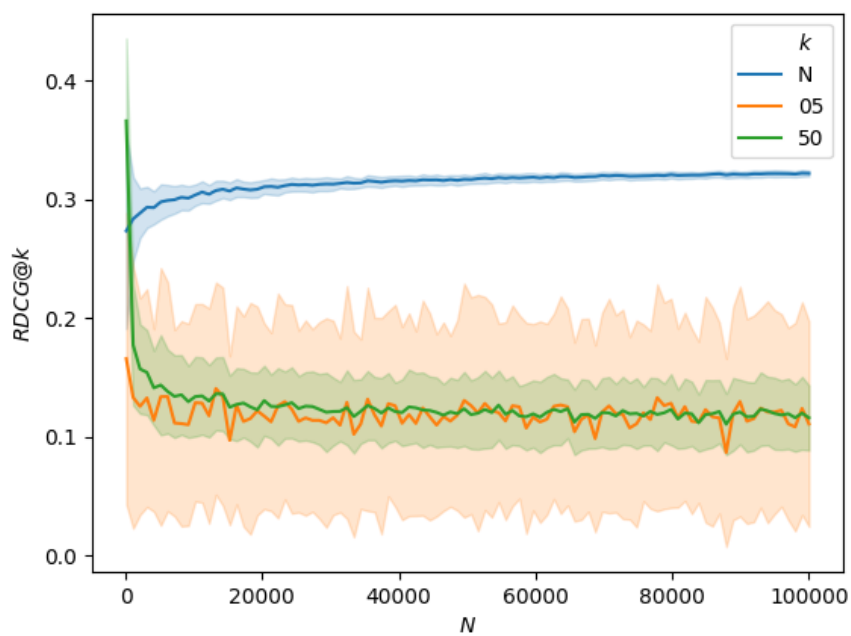


Figura 7 – $RDCG$ e $RDCG@k$ em função de N até 10^4

fácilmente interpretáveis, como o esperado conforme sua formulação. A métrica parece convergir para um valor < 1 ao utilizar f constante, e já para N nas ordens de grandeza abaixo de 10^3 está estabilizar por volta de 0.31.

Figura 8 – $RDCG$ e $RDCG@k$ em função de N até 10^5

Considerações finais

O $NDCG$ tem sua propriedade de convergência exibida experimentalmente, após demonstrada sua convergência quase certa no caso geral de múltiplos níveis de relevância. Além disso, também se exibiu o comportamento divergente de múltiplas métricas de ranqueamento limitadas aos top- k elementos, evidenciando que são métricas com características numéricas indesejadas - valendo para $P@k$, $\bar{P}@k$, $NDCG@k$ e $RDCG@k$.

Através do experimento também foi possível verificar pontos de atenção no uso do $NDCG$, como sua rápida convergência e a consequente dificuldade de utilizar seus valores absolutos para comparar performances de sistemas de ranqueamento quando o tamanho do ranking não é constante.

Por fim, foi possível observar que o $RDCG$ experimentalmente apresentou boas características numéricas, em particular a aparente convergência para um valor < 1 e sua estabilidade em valores pequenos ($0.3 \sim 0.4$) para rankings nos cenários de busca simulados. A demonstração teórica da convergência e distinguibilidade consiste do $RDCG$ é uma possível direção de trabalhos futuros.

Os resultados dos experimentos indicam que para o cenário de maior interesse neste trabalho (proporcionalidade de documentos relevantes, ou distribuição de relevâncias constante para rankings de tamanhos diferentes), métricas clássicas como \bar{P} podem ser mais úteis do que o $NDCG$. Alternativamente, o $RDCG$ também apresenta as propriedades numéricas desejadas e mantém o modelo conceitual de ganho cumulativo descontado do DCG .

O comportamento convergente foi observado em conjuntos de até 10 mil documentos, o que é representativo dos casos de uso mais frequentemente modelados no cenário objetivo deste trabalho: triagem gradativa de documentos em um processo de busca iterativo. Por exemplo, o cenário citado de processos seletivos de currículos costuma contar com a ordem de grandeza de dezenas até milhares de currículos inscritos em uma vaga.

Uma limitação dos experimentos realizados é considerar apenas funções avaliadores constantes, sendo sistemas de recuperação de informação irreais que não fazem nenhuma espécie de ranqueamento. É possível que ao utilizar funções avaliadoras verdadeiras, \bar{P} ou $RDCG$ apresentem outros tipos de comportamentos numéricos indesejados que façam o $NDCG$ ser a melhor opção. Apesar disso, comparar um sistema de ranqueamento com um sistema simplificado (para *benchmarking*) justifica a preferência por métricas que sirvam para o caso de funções avaliadoras constantes.

A métrica $RDCG$ implementa apenas em parte a sugestão presente em ([KATE-](#)

RENCHUK; ROSENBERG, 2018) chamada *rankDCG*, que não somente normaliza o *DCG* de forma diferente, mas também remapeia os graus de relevância para um conjunto de graus de relevância que reflete apenas a ordenação (e não graus de relevância relativa). Tal intervenção nos graus de relevância possivelmente afeta a convergência das métricas de ranqueamento, além da distinguibilidade consistente. Analisar o comportamento assintótico do *rankDCG* também pode ser uma expansão futura deste trabalho.

De forma conclusiva, ficou demonstrado que a métrica mais comumente utilizada em recuperação da informação (*NDCG*) pode ser substituída por outra métrica (\bar{P} ou *RDCG*) para facilitar a interpretação de resultados e aumentar a confiabilidade no diagnóstico da qualidade de funções avaliadoras em sistemas de ranqueamento que lidam com conjuntos crescentes de documentos relevantes. Tais cenários, apesar de menos prevalentes em estudos de métricas de ranqueamento, possuem aplicações reais e características que demandam diferentes expectativas de comportamento assintótico das métricas.

Referências

- BURGES, C. et al. Learning to rank using gradient descent. In: *Proceedings of the 22nd international conference on Machine learning - ICML '05*. ACM Press, 2005. p. 89–96. ISBN 1595931805. ISSN 00243205. Disponível em: <<http://dl.acm.org/citation.cfm?id=1102351.1102363>><http://portal.acm.org/citation.cfm?doid=1102351.1102363>>. Citado na página 19.
- COOPER, W. S. Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation*, 1968. ISSN 1936-6108. Citado na página 11.
- JÄRVELIN, K.; KEKÄLÄINEN, J. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, v. 20, n. 4, p. 422–446, 2002. ISSN 10468188. Disponível em: <<http://portal.acm.org/citation.cfm?doid=582415.582418>>. Citado 2 vezes nas páginas 12 e 19.
- KATERENCHUK, D.; ROSENBERG, A. RankDCG: Rank-Ordering Evaluation Measure. 2018. Disponível em: <<http://arxiv.org/abs/1803.00719>>. Citado 2 vezes nas páginas 24 e 36.
- LIU, T.-Y. *Learning to Rank for Information Retrieval*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. ISBN 978-3-642-14266-6. Disponível em: <<http://link.springer.com/10.1007/978-3-642-14267-3>>. Citado 4 vezes nas páginas 12, 17, 18 e 19.
- MANNING, C. D.; RAGHAVAN, P.; SCHUTZE, H. *Introduction to Information Retrieval*. [S.l.: s.n.], 2008. Citado na página 11.
- PRADEEP, R.; TEWARI, A.; YANG, E. On NDCG consistency of listwise ranking methods. *Journal of Machine Learning Research*, v. 15, p. 618–626, 2011. ISSN 15324435. Citado na página 12.
- ROBERTSON, S.; BELKIN, N. RANKING IN PRINCIPLE. *Journal of Documentation*, v. 34, n. 2, p. 93–100, feb 1978. ISSN 0022-0418. Disponível em: <<http://www.emeraldinsight.com/doi/10.1108/eb026654>>. Citado 2 vezes nas páginas 12 e 19.
- WANG, J. et al. IRGAN: A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models. v. 10, 2017. ISSN 16130073. Disponível em: <<https://arxiv.org/pdf/1705.10513.pdf>>. Citado na página 19.
- WANG, Y. et al. A Theoretical Analysis of NDCG Type Ranking Measures. apr 2013. ISSN 1938-7228. Disponível em: <<http://arxiv.org/abs/1304.6480>>. Citado 4 vezes nas páginas 17, 19, 21 e 24.