Análise de Eficácia baseada em Diversidade para Sistemas de Recuperação de Informação

Raul do Carmo Peixoto, José Solenir L. Figuerêdo, Rodrigo Tripodi Calumby

[1] Universidade Estadual de Feira de Santana – Feira de Santana – BA – Brazil

raul.p.carmo@gmail.com, jslfigueredo@ecomp.uefs.br, rtcalumby@uefs.br

Abstract. *Information retrieval systems are generally evaluated considering ranking relevance only. However, this does not always allow us to assess how good the result was, considering the user's intent. Thus, it is necessary to incorporate diversity measures into the assessment process. In this work we present extensions incorporated to AnalyzIR, which is a practical tool for analyzing the effectiveness of information retrieval systems. By default, the tool only provided relevance analysis, but by aggregating frameworks and diversity measures, it became an integrated and intuitive platform that considers both the relevance and the diversity of the results.*

Resumo. *Os sistemas de recuperação de informação geralmente são avaliados considerando apenas a relevância do ranking. Porém, isso nem sempre nos permite avaliar o quão bom foi o resultado, considerando a intenção do usuário. Assim, é necessário incorporar medidas de diversidade ao processo de avaliação. Neste trabalho apresentamos extensões incorporadas à AnalyzIR, que é uma ferramenta prática de análise de eficácia para sistemas de recuperação de informações. Por padrão, a ferramenta disponibilizava apenas análises por relevância, mas ao agregar frameworks e medidas de diversidade, tornou-se uma plataforma integrada e intuitiva, que considera tanto a relevância quanto a diversidade dos resultados.*

1. Introdução

Sistemas de Recuperação da Informação (RI) visam a obtenção de dados requisitados por um usuário por meio de uma consulta em uma base de dados. O resultado da busca geralmente é uma lista de itens que visam satisfazer a requisição feita, e que são ranqueados seguindo algum critério de relevância. Comumente, esse ranqueamento é realizado com base na similaridade dos itens em relação à consulta. Contudo, considerando as complexas tarefas de busca impostas aos sistemas de RI, especialmente sistemas para busca de imagens e vídeos, o uso exclusivo da similaridade como critério pode resultar em dois problemas principais: redundância e ambiguidade [Rodrygo et al. 2015]. A redundância surge uma vez que diferentes itens, apesar de igualmente relevantes, podem trazer dados equivalentes, não acarretando em ganho de informação para o usuário. Por sua vez, a ambiguidade decorre de possíveis interpretações da real intenção de busca do usuário expressada por meio da consulta.

Para atenuar os problemas mencionados, diversos trabalhos têm proposto a introdução de procedimentos de diversificação de resultados nos sistemas de RI [Xu et al. 2017]. Com isso, surge a necessidade de quantificar a eficácia desses

sistemas segundo a diversidade. Para isso, várias medidas e *frameworks* foram criados: Alpha [Clarke et al. 2008], IA [Agrawal et al. 2009], D e D# [Sakai et al. 2010], Subtopic-Recall e Subtopic-Precision [Zhai et al. 2015]. Contudo, para o nosso conhecimento, ainda não existe uma ferramenta que integre esses recursos de análise de eficácia. Diante disso, desenvolvemos a ferramenta AnalyzIR, a qual faz a integração dos frameworks e medidas de análise de diversidade à ferramenta AnalyzIR. Esta ferramenta está sendo desenvolvida na UEFS e já oferece, entre outras funcionalidades, a análise de relevância e significância estatística, sendo uma plataforma integrada que combina computo de medidas de eficácia, representações gráficas e comparação de resultados.

Para avaliação de resultados, existem medidas que representam o nível de diversidade de modo direto ou que são construídas a partir de *frameworks*. Os *frameworks* funcionam utilizando medidas de relevância tradicionais e adaptando-as para quantificar a diversidade do resultado da consulta. Neste trabalho realizou-se o levantamento dos *frameworks* e medidas de diversidade disponíveis na literatura. Os frameworks selecionados para implementação na ferramenta foram o IA, D e D#. O IA foi implementado para utilizar as medidas tradicionais de relevância já disponíveis na AnalyzIR, transformando-as em medidas de diversidade. Adicionalmente, a ferramenta NTCIR Eval [Sakai 2011] foi integrada à ferramenta base a fim de trazer medidas específicas de diversidade e implementar os frameworks D e D#. Dentre estas medidas, foi incluída a Subtopic-Recall. Além dos frameworks citados, também for incorporada a medida F_1 para computo da média harmônica, comumente aplicada para integrar medidas de relevância e diversidade. As funcionalidades foram testadas a partir de um conjunto de dados sintéticos.

2. Frameworks e Medidas de Avaliação de Diversidade

2.1. Alpha (α-DCG)

O α-DCG (Eq. 1) utiliza a redundância que pode ser gerada por uma consulta para aplicar penalização, denotada pelo termo $(1-\alpha)^{\sum_{t=1}^{k} J(d_t,n_i)}$ em que α representa quão rigorosa será. Assim, quanto maior o valor de α, menor será o valor de $(1-\alpha)$. Essa parcela é elevada ao termo $\sum_{t=1}^{j-1} J(d_t, n_i)$, onde $J(d_t, n_i)$ representa o julgamento binário (i.e., 1 caso o item seja relevante, ou 0 caso contrário) dos documentos d na posição t em relação àquela intenção (n_i). O α-DCG penaliza duas vezes um documento em um ranking, uma penalização baseada na relevância e outra na redundância, $\log_2(j+1)$ e $(1-\alpha)^{\sum_{t=1}^{j-1} J(d_t,n_i)}$, respectivamente.

$$\alpha\text{-}DCG(k) = \sum_{j=1}^{k} \frac{\sum_{i=1}^{m} J(d_j, n_i)(1-\alpha)^{\sum_{t=1}^{j-1} J(d_t,n_i)}}{\log_2(j+1)} \tag{1}$$

2.2. Framework Intent-Aware (IA)

O *Framework* IA (Eq. 2) baseia-se no conhecimento das possíveis intenções associadas à consulta e em estimativas de suas probabilidades ($P(a|q)$). Por exemplo, tomando-se que uma consulta pela palavra "Bond" levou 70% dos usuários a clicar em um artigo acerca da série de filmes "James Bond" e os outros 30% restantes clicaram em outros tópicos, isso significa que a probabilidade dessa consulta ter como intenção os filmes é de $0,70$.

O termo $Eval(Q, k)$ representa um framework de relevância, onde Q é um conjunto de itens, e k denota quantas intenções foram consideradas para a consulta.

$$IA\text{-}Eval(Q, k) = \sum_{i=1}^{k} P(a|q)Eval(Q, k) \tag{2}$$

2.3. Framework D e D#

O Framework D# recompensa ranqueamentos que satisfaçam duas condições: i) Que o *rank* contenha itens associados ao maior número de intenções possíveis e ii) Que itens associados a intenções populares sejam ranqueados acima dos itens associados a intenções marginais. Para atendar à premissa i), usa-se $S - recall@k$, tendo em vista que essa medida recompensa ranqueamentos com alta cobertura de intenções. Por sua vez, para a premissa ii), pode ser utilizada qualquer medida que compute relevância relativa às intenções, como a $GG(r)$, representada pela Eq. 3.

$$GG(r) = \sum_{i \in I_q} P(i|q)g_i(d) \tag{3}$$

Onde I_q é o conjunto de intenções de uma consulta q; O termo $g_i(d)$ denota os ganhos dos documentos em relação à intenção i; e $P(i|q)$ é a probabilidade da intenção i em relação à consulta q. Similarmente ao DCG, definimos também um $GG'(r)$ para uma consulta ideal, isto é, uma consulta onde os documentos são ranqueados de maneira decrescente de acordo com sua $GG(r)$; Assim, podemos normalizar o valor de $GG(r)$ para obter o $div\text{-}nDCG(r)$, conforme Eq. 4. Por fim, é definido o $Idiv\text{-}nDCG@r$ (Eq. 5) como a combinação linear entre o $S\text{-}recall@r$ e o $div\text{-}nDCG@r$.

$$div\text{-}nDCG(r) = \frac{\sum_{i=1}^{r} \frac{GG(i)}{\log(i+1)}}{\sum_{i=1}^{r} \frac{GG'(i)}{\log(i+1)}} \tag{4}$$

$$Idiv\text{-}nDCG@r = \gamma S\text{-}recall@r + (1 - \gamma)div\text{-}nDCG@r \tag{5}$$

2.4. Subtopic Recall e Suptopic Precision

O *Subtopic Recall* ou *S-recall@k* avalia a eficácia de um sistema de RI baseado na quantidade de aspectos ou "*Informational Nuggets*" cobertos pelos documentos recuperados. Formalmente, dado um resultado de tamanho k, e uma consulta q, e A_q como o conjunto de aspectos subentendidos para consulta q, defini-se $S\text{-}recall@k$ como o percentual de aspectos da consulta cobertos pelos itens retornados, conforme Eq. 6. $|A_q|$ é a quantidade de aspectos subentendidos pela consulta, e a função $subtopics(d_i)$ retorna a quantidade de aspectos cobertos pelo item *d* da posição *i* do resultado.

$$S\text{-}recall@k = \frac{|\bigcup_{i=1}^{k} subtopics(d_i)|}{|A_q|} \tag{6}$$

A *Subtopic Precision (S-precision@r)*, medida complementar à $S\text{-}recall@k$, permite indicar com quantos itens pode-se alcançar um determinado valor de $S\text{-}recall@k$.

Formalmente, $S\text{-}precision@r$ (Eq. 7) é definido como a quantidade de itens retornados por um sistema, dividido pela quantidade de itens retornados pelo sistema para um dado $S\text{-}recall@k$. Ou seja, dado um $S\text{-}recall@k$, o $S\text{-}precision@r$ indica quão bem os documentos cobrem as intenções da consulta. Caso um sistema obtenha um $S\text{-}precision@r$ baixo, isso significa que os documentos retornados, individualmente, cobrem poucas intenções.

$$S\text{-}precision@r = \frac{|minRank(S_{opt}, r)|}{|minRank(S, r)|} \tag{7}$$

3. Resultados e discussão

As implementações aqui descritas foram realizadas sobre a ferramenta AnalyzIR, que permite ao usuário pode criar representações gráficas significativas e úteis a partir de seus dados de avaliação. Este trabalho permitiu estender a ferramenta para análises de sistemas baseados em diversidade, conforme descrevemos a seguir. Para fazer a análise de um sistema, o usuário deve criar um Projeto na categoria diversidade (Figura 1). Para isso, ele deve fornecer um nome para o projeto, um diretório de trabalho onde os dados serão armazenados e utilizados durante os cálculos, além do arquivo "Qrels" (ground-truth), um arquivo "Cluster Qrels" (ground-truth de diversidade), um arquivo opcional "IProb" (probabilidades de cada intenção de busca) e os resultados que serão analisados ("Runs").

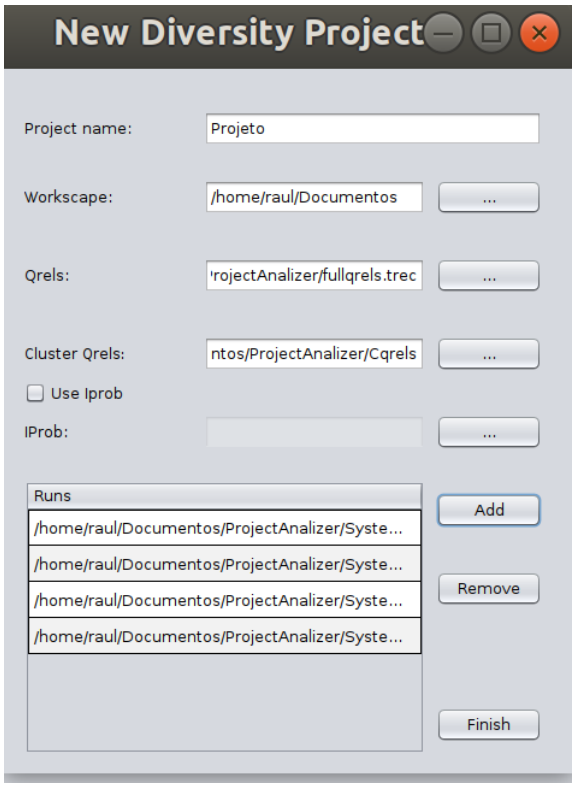

Figura 1. Criação do projeto de avaliação da diversidade.

Com o projeto criado, o usuário é capaz de criar gráficos, definindo os *frameworks* de interesse, quais resultados serão analisados e os parâmetros de cálculo da medida selecionada. Para criar uma análise visual, inicialmente deve-se escolher um *framework* de diversidade (Figura 2). O próximo passo é selecionar os dados a serem utilizados na criação

do gráfico (Figura 3): resultados (“runs”); consultas de interesse (“topics”); e as medidas de diversidade do framework escolhido. Dependendo da escolha feita pelo usuário, ele precisará fornecer valores para alguns parâmetros, como é o caso do framework D# ou Alpha, onde o usuário deverá indicar os valores de Gamma e Alpha.

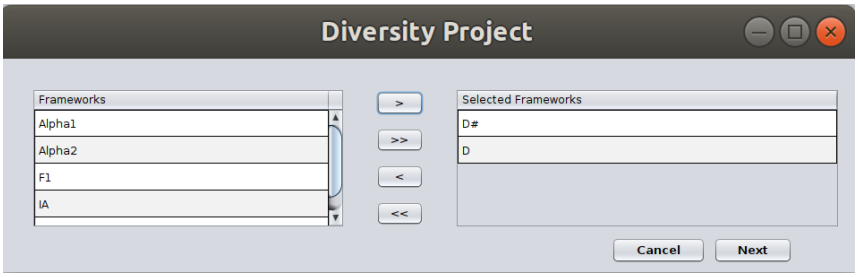

Figura 2. Definição do *framework* de interesse.

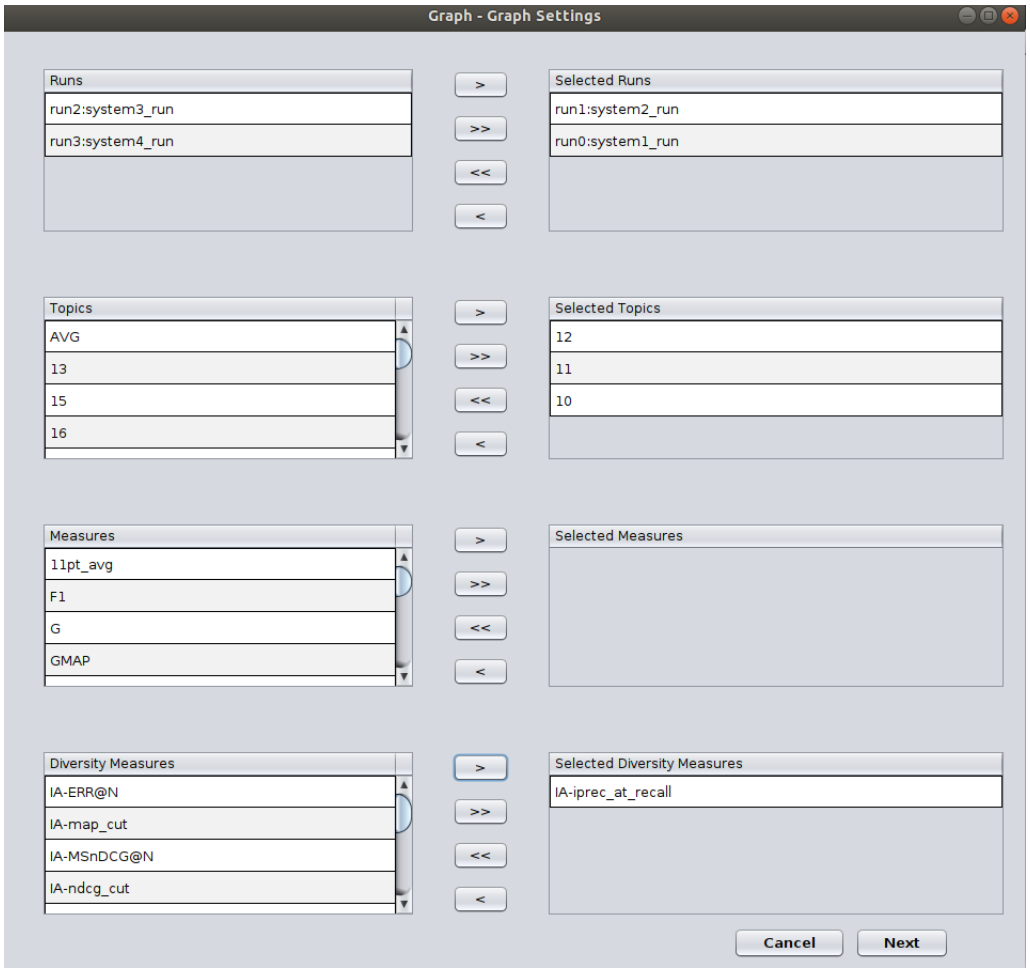

Figura 3. Seleção de dados para criação de gráfico.

A Figura 4 apresenta um exemplo de gráfico utilizando o sistema “run1”, as consultas 10, 11 e 12 e a medida $IA\text{-}P@N$. Além da visualização dos resultados no gráfico em si, o usuário possui algumas opções adicionais, podendo editar, exportar, abrir em uma nova janela e editar os rótulos. Além disso, é possível alterar o tópico que está sob análise ou considerar o resultado médio para múltiplas consultas.

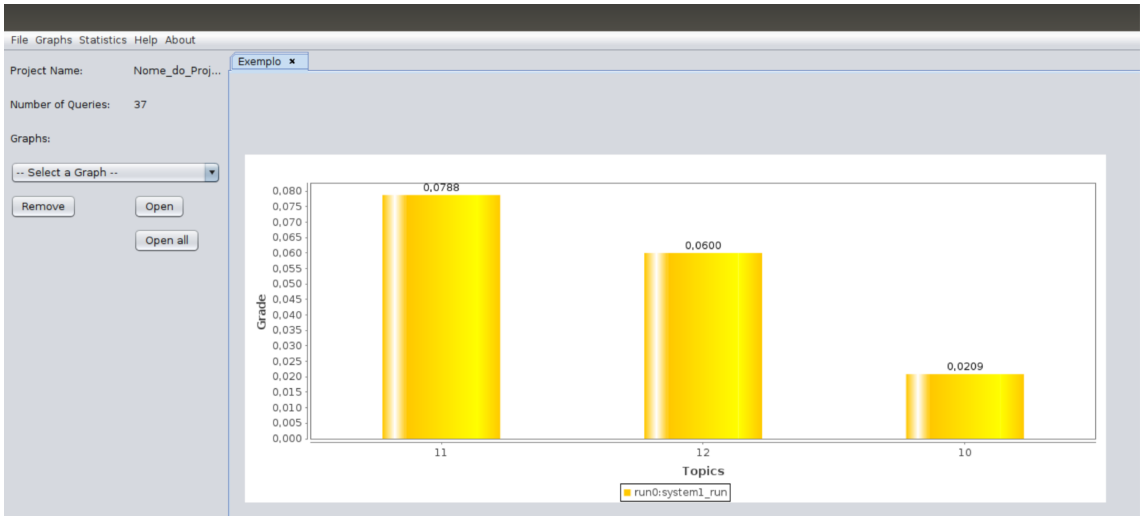

Figura 4. Exemplo de resultado com o sistema *run1* para os tópicos 10, 11 e 12 e a medida *IA-P@N*.

4. Conclusão

Ao adicionar medidas de diversidade, a AnalyzIR tornou-se ainda mais robusta, oferecendo uma plataforma integrada para análise de diversidade em resultados de sistemas de recuperação da informação. Essas funcionalidades, em conjunto com as demais já disponíveis, oferecem uma plataforma única e intuitiva para pesquisadores e profissionais da área de RI. Como trabalhos futuros, planejamos melhorar a interface gráfica e adicionar novas medidas/*frameworks* de diversidade. Além disso, pretende-se migrar a ferramenta para o ambiente web.

Referências

Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. (2009). Diversifying search results. In *International Conference on WSDM*, page 5–14, New York, NY, USA.

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *ACM SIGIR*, page 659–666, New York, NY, USA.

Rodrygo, L. S., Macdonald, C., and Ounis, I. (2015). Search result diversification. *Foundations and Trends in Information Retrieval*, 9(1):1–90.

Sakai, T. (2011). Ntcireval: A generic toolkit for information access evaluation. In *Proceedings of the forum on information technology*, volume 2, pages 23–30.

Sakai, T., Craswell, N., Song, R., Robertson, S., Dou, Z., and yew Lin, C. (2010). Simple evaluation metrics for diversified search results. In *EVIA*, pages 42–50.

Xu, J., Xia, L., Lan, Y., Guo, J., and Cheng, X. (2017). Directly optimize diversity evaluation measures: A new approach to search result diversification. *ACM Transactions on Intelligent Systems and Technology*, 8(3).

Zhai, C., Cohen, W. W., and Lafferty, J. D. (2015). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. *SIGIR Forum*, 49(1):2–9.