UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto nº17.228 de 25/11/2016
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXIV SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2020

Estudo e Desenvolvimento de Ferramentas de Análise Estatística para Sistemas de Recuperação de Informação Interativa

Moisés Almeida da Cruz Farias[1] e Rodrigo Tripodi Calumby[2]
1. Bolsista PIBIC/CNPq, Graduando em Eng. da Computação, UEFS, e-mail: moisesalmeida123_@hotmail.com
2. Orientador, Departamento de Ciências Exatas, UEFS, e-mail: rtcalumby@uefs.br

PALAVRAS-CHAVE: Recuperação de informação, sistemas interativos, análise estatística.

INTRODUÇÃO

Atualmente, há várias ferramentas de captura e armazenamento de informações que são utilizadas diariamente por grande parte da população. Uma das consequências disso é a produção de uma grande massa de dados digitais que podem ser usadas em diferentes áreas do conhecimento. Assim, é essencial aplicar métodos eficazes de recuperação da informação para que esses dados possam ser utilizados da melhor maneira possível (Calumby *et al.,* 2016). Os métodos que utilizam alguma forma de interação com o usuário, por exemplo, via feedback do usuário para indicar se uma informação é relevante ou não, são chamados de sistemas interativos. Cada ciclo de interação do usuário com o resultado de uma busca define uma iteração (Järvelin *et al*., 2008); (Kanoulas *et al*., 2011). Para mensurar a eficácia, comparar sistemas, construir análises gráficas e realizar testes estatísticos para este tipo de sistema é necessária a utilização de um conjunto diverso de ferramentas, de modo trabalhoso e suscetível a erros e incompatibilidades. Para atenuar estes desafios, está em desenvolvimento na UEFS a *AnalyzIR*, uma ferramenta para apoio na análise de eficácia de sistemas de recuperação da informação e que integra diversas das funcionalidades necessárias. Este trabalho consistiu-se do projeto, desenvolvimento e integração de duas funcionalidades já existentes na ferramenta: a avaliação de sistemas interativos e a realização de testes de significância estatística. Estes testes são feitos comparando os resultados das iterações de sistemas de interesse com outro sistema de referência (baseline) utilizando métodos, sendo *Wilcoxon's Signed Rank Test*, *Mann-Whitney U* e *teste t de Student*.

METODOLOGIA

Para o desenvolvimento das novas funcionalidades, foi realizado o estudo bibliográfico sobre a avaliação de sistemas de recuperação interativa da informação, testes de significância estatística e a utilização destes para a análise de sistemas interativos. Além disso, a estrutura interna da AnalyzIR foi estudada para determinar-se a melhor forma de implementar as novas funcionalidades e possíveis necessidades de refatoração das

funcionalidades já disponíveis. Para gerenciar a integração das novas funcionalidades à ferramenta base foi utilizado um repositório online para o código fonte, baseado no sistema git. Para codificação, depuração e experimentação foram utilizados os softwares Eclipse IDE e IntelliJ IDEA.

RESULTADOS E DISCUSSÃO

Ao realizar uma avaliação de sistemas interativos na AnalyzIR, o usuário pode escolher se quer ou não utilizar os testes estatísticos, sendo em caso positivo o resultado apresentado junto com a avaliação e gráficos do sistema interativo. Os passos seguintes levam em consideração que foi criado um projeto para análise de sistemas interativos, que a opção de criação de gráficos foi selecionada e que os sistemas a serem comparados (runs), iterações, consultas (buscas do usuário) e medidas (métodos que mensuram a eficácia de um sistema) já foram definidas pelo usuário.

Após a escolha da configuração do gráfico, o usuário é direcionado para a tela de configuração do teste estatístico (Figura 1), onde seleciona o tipo de teste; o baseline; e o nível de significância desejado (alpha). Caso o usuário escolha o Wilcoxon ou o Mann-Whitney *U*, o nível de significância poderá ser escolhido livremente, mas no caso do teste t de Student, os níveis serão disponibilizados pela ferramenta e o usuário terá que escolher um deles, dada a utilização da tabela de valores críticos que considera níveis pré-definidos de significância e grau de liberdade; e, por fim, seleciona-se a biblioteca que disponibiliza a implementação do teste estatístico.

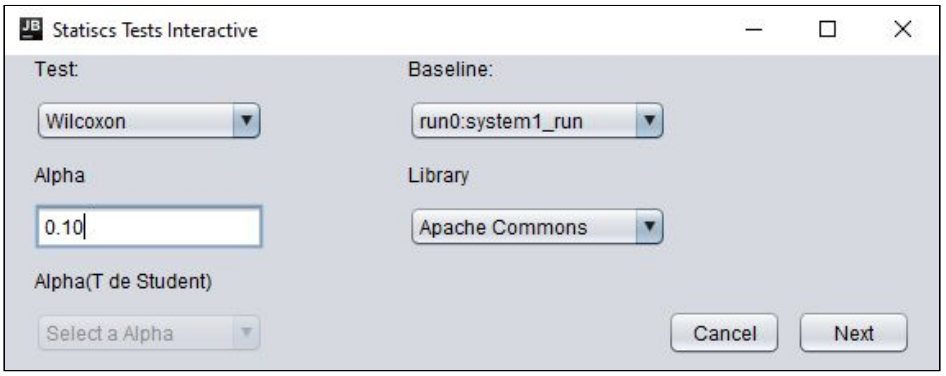

Figura 1: Tela de configuração do teste estatístico. No exemplo foi selecionado o teste de Wilcoxon. O baseline é o system1, a biblioteca é a Apache Commons e o nível de significância é de 0,10.

O gráfico será apresentado ao usuário contendo o resultado da avaliação dos sistemas e o resultado do teste estatístico. Os dados apresentados são a média dos resultados das consultas, mas para o cálculo de significância estatística é utilizado o resultado da avaliação por consulta (modo pareado). O gráfico gerado pode ser de curva ou de barra, dependendo da quantidade de iterações escolhidas pelo usuário.

No gráficos de barra, para representar se existe ou não significância estatística foi utilizado o próprio valor encontrado na avaliação do sistema. Caso exista significância, a cor será azul para superioridade ou vermelha para inferioridade. Se não há significância, mantém-se em preto. Além disso, na legenda haverá conter um "*" no(s) sistema(s) para os quais há significância estatística, como ilustrado na Figura 2.

No gráfico da Figura 2, foi analisada apenas uma iteração dos sistemas e o baseline escolhida foi o system1. O resultado indica que houve significância para os sistemas system2 e system4, indicado na legendo com "*" e pelo valor em vermelho na barra, indicando inferioridade.

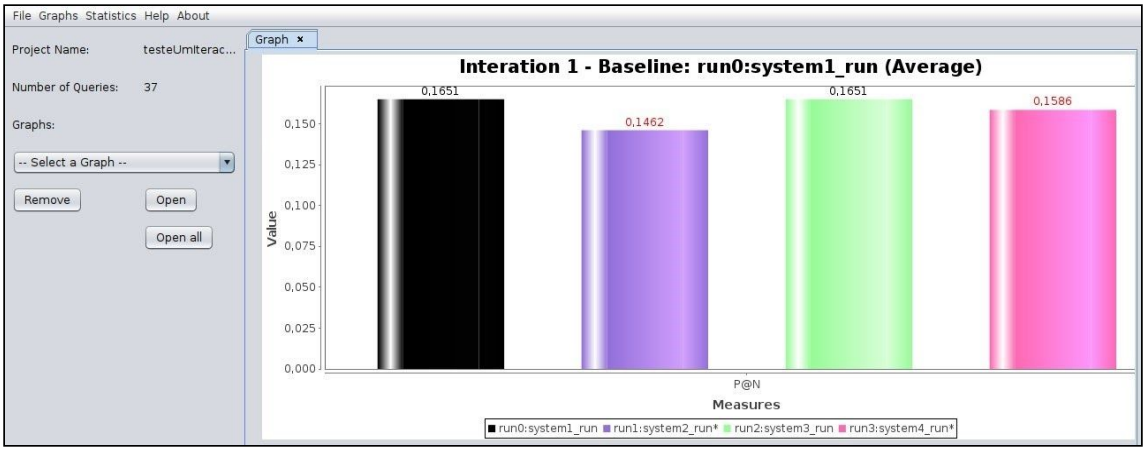

Figura 2: Exemplo de gráfico do tipo barra. Resultado da avaliação de eficácia e da análise de significância estatística de quatro sistemas com apenas uma iteração.

Nos gráficos de curva (Figura 3), para representar se houve ou não significância estatística foram utilizados os pontos nas curvas. Caso não exista significância, a cor do ponto será igual ao da linha, mas caso exista, utilizará o mesmo sistema de cores (azul ou vermelho) descrito anteriormente.

No gráfico da Figura 3 foram analisadas quatro iterações, sendo o system9 o baseline. De acordo com o resultado observa-se que na segunda iteração houve significância inferior (indicado pelo ponto em vermelho) nos sistemas system6 e system7. Também houve significância nas demais iterações dos dois sistemas citados anteriormente e todas apresentam superioridade (indicado pelo ponto em azul).

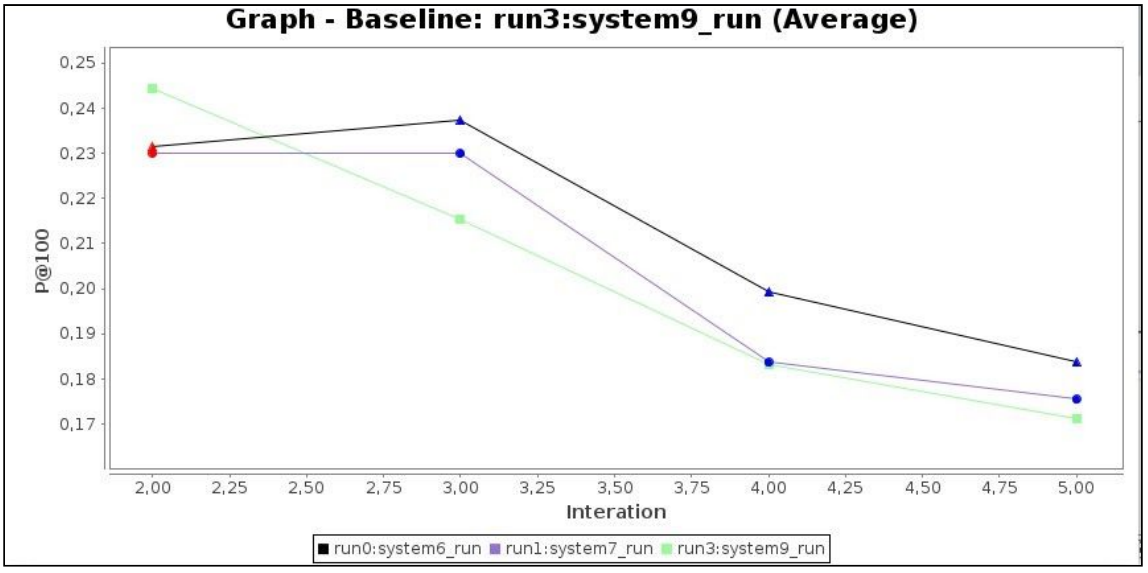

Figura 3: Exemplo de gráfico do tipo curva. Resultado da avaliação de eficácia e da análise de significância estatística de três sistemas ao longo de quatro iterações. Indicando significância nas quatro iterações dos sistemas comparados.

Depois do gráfico criado é possível exportá-lo em diferentes formatos: jpg, png, xls e csv. O formato xls contém os dados do resultado da análise do sistema (run) e do

teste estatístico, como é possível observar na Figura 4. A primeira coluna contém os nomes dos sistemas; A segunda o número da iteração; A terceira coluna é o resultado que a run obteve na avaliação e a medida escolhida pelo usuário; Na quarta coluna o valor obtido no teste estatístico onde a primeira linha indica o nome do teste escolhido; A quinta coluna indica se houve significância estatística, "-" indica que não houve e "significant" significa que houve; e no cabeçalho da quinta coluna há o nível de significância escolhido; A sexta coluna indica o baseline utilizado.

	A	B	C	D	E	F	G
1	Run	Iteration	P@N	p_value-wilcoxon	Significant-0.9	Baseline	
2	run0:system6_run	4	0.4333	0.045500263896358306	-	run0:system6_run	
3	run0:system6_run	3	0.5091	0.0455002638963583	-	run0:system6_run	
4	run0:system6_run	2	0.4	0.0455002638963583	-	run0:system6_run	
5	run1:system7_run	3	0.4833	0.045500263896358306	-	run0:system6_run	
6	run1:system7_run	4	0.3812	0.0455002638963583	-	run0:system6_run	
7	run1:system7_run	2	0.45	0.0455002638963583	-	run0:system6_run	
8	run2:system8_run	4	0.4333	0.045500263896358306	-	run0:system6_run	
9	run2:system8_run	3	0.5091	0.0455002638963583	-	run0:system6_run	
10	run2:system8_run	2	0.4	0.0455002638963583	-	run0:system6_run	
11	run3:system9_run	2	0.5286	0.045500263896358306	-	run0:system6_run	
12	run3:system9_run	3	0.5	0.0455002638963583	-	run0:system6_run	
13	run3:system9_run	4	0.4267	0.0455002638963583	-	run0:system6_run	

Figura 4: Exemplo de arquivo exportado para o formato xls que contém as informações da avaliação dos sistemas e o resultado da análise estatística.

CONSIDERAÇÕES FINAIS

Após o projeto finalizado, foi possível unir duas funções que já estavam implementadas na ferramenta: a avaliação de sistemas interativos e os testes estatísticos. Atualmente é possível que o usuário possa aplicar o teste em sistemas interativos e ter um retorno visual tanto do resultado da análise do sistema como do teste. Além disso, é possível exportar os dados de avaliação do sistema e os do teste estatístico para serem estudados posteriormente, assim, se tornando uma ferramenta mais completa para o uso dos pesquisadores. O conhecimento adquirido com esse projeto foi muito importante e crucial para entender a importância da avaliação estatística para validar os resultados de um experimento.

REFERÊNCIAS

CALUMBY, R. T.; GONÇALVES, M. A.; TORRES, R. da S. 2016. On Interactive Learning-to-Rank for IR: Overview, Recent Advances, Challenges, and Directions. Neurocomputing, Amsterdam, 208:3-24.

JÄRVELIN, K.; PRICE, S. L.; DELCAMBRE, L. M. L.; NIELSEN, M. L. 2008. Discounted cumulated gain based evaluation of multiple-query ir sessions, In: Proceedings of 30th European Conference on Advances in Information Retrieval, Springer-Verlag, Berlin, Heidelberg, p. 4–15.

KANOULAS, E.; CARTERETTE, B.; CLOUGH, P. D.; SANDERSON, M. 2011. Evaluating multi-query sessions, In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, pp. 1053–1062.