Ferramenta de Análise Estatística para Sistemas de Recuperação de Informação Interativa

Moisés Almeida da Cruz Farias[1], Rodrigo Tripodi Calumby[2]

[1]Engenharia da Computação,Universidade Estadual de Feira de Santana(UEFS)
Feira de Santana - BA - Brasil
[2]Departamento de Ciências Exatas,Universidade Estadual de Feira de Santana(UEFS)
Feira de Santana - BA - Brasil

macfarias@ecomp.uefs.br, rtcalumby@uefs.br

abstract. *The rigorous analysis of the effectiveness of information retrieval systems is a process that requires the use of complex techniques and multiple tools. Assessing the effectiveness of search engines brings challenges considering the iterative process performed by users. The AnalyzIR tool is being developed to provide integrated functionality for analyzing the effectiveness of information retrieval systems. This work presents an extension of the tool designed to support the statistical significance analysis of the effectiveness of interactive retrieval systems. For this, methods of computing effectiveness measures and analysis of statistical significance for interactive systems were integrated.*

Resumo. *A análise rigorosa da eficácia de sistemas de recuperação da informação é um processo que demanda o uso de técnicas complexas e múltiplas ferramentas. Avaliar a eficácia dos sistemas em sessões de busca traz desafios adicionais considerando o processo iterativo realizado pelos usuários. A ferramenta AnalyzIR está sendo desenvolvida para fornecer funcionalidades integradas para análise da eficácia de sistemas de recuperação da informação. Este trabalho apresenta a extensão da ferramenta projetada para apoio à análise de significância estatística da eficácia de sistemas de recuperação interativa. Para isso, foram integrados métodos de cômputo de medidas de eficácia e análise de significância estatística para sistemas interativos.*

1. Introdução

Dado o avanço e popularização das tecnologias de captura e armazenamento de informação, um conjunto grande de conteúdos digitais tem sido criado. A necessidade de explorar de forma eficiente estas bases de dados é apresentada em várias áreas, como: ciências biológicas e da natureza, medicina, redes sociais, recomendação ao usuário, etc. Desta forma, a construção de métodos eficazes para recuperação e indexação destas bases é imprescindível [Calumby et al. 2016].

Conteúdos do tipo imagem, permitem dois tipos básicos de abordagens para recuperação: baseada em texto ou baseada em conteúdo. Em comparação a abordagem

baseada em texto, a baseada no conteúdo da imagem [Torres and Falcão 2006] (CBIR, do inglês, *Content-based Image Retrieval*) possibilita analisar outras características e então estimar a relevância segundo outros critérios, como cores e texturas. A escolha adequada de descritores visuais é imprescindível para métodos de CBIR. Descritores visuais são responsáveis por definir quão similares são duas imagens a partir dos algoritmos de extração de características e de distância [Penatti 2009]. O resultado obtido por esses descritores vai definir o grau de relevância de cada imagem da base. A abordagem da técnica de realimentação de relevância é possibilitar ao usuário exprimir a sua necessidade sem ter que conhecer propriedades de baixo nível das imagens. Esse processo é realizado iterativamente e interativamente [Calumby 2010]. A cada iteração, o algoritmo de realimentação de relevância busca capturar quais propriedades visuais melhor definem as imagens informadas como relevantes pelo usuário.

Atualmente, para efetuar o comparativo de sistemas usando medidas de eficácia, pesquisadores e profissionais da indústria necessitam utilizar um conjunto de ferramentas para mensurar a eficácia, realizar análise estatística destes valores, fazer análise de correlação para diferentes cenários e construir modelos visuais para apresentação destes resultados. Em especial, para garantir o rigor científico na análise de eficácia de diferentes métodos, é necessária a utilização de testes para determinação da significância estatística considerando a execução dos métodos em múltiplas bases de dados e múltiplas consultas. Devido a utilização dessas diversas ferramentas a ocorrência de erros se torna mais suscetível, e além disso ainda existe o possível obstáculo das extensões dos arquivos resultantes de cada ferramenta ser diferente.

Por isso, a ferramenta AnalyzIR tem sido desenvolvida na UEFS. Essa ferramenta visa disponibilizar ao usuário um ambiente integrado de avaliação dos resultados de métodos de recuperação de informação, com diversas funcionalidades como por exemplo: a avaliação de sistemas não interativos e interativos, testes estatísticos e análise de correlação para sistemas não interativos, entre outros. O objetivo do trabalho é integrar à ferramenta AnalyzIR, a funcionalidade de análise estatística para sistemas interativos de recuperação da informação, assim possibilitando a comparação de um conjunto desses sistemas em relação há um determinado sistema baseline, para identificar se existe significância estatística entre dois sistemas.

2. Materiais e Métodos

Considerando a ferramenta base deste trabalho, as novas funcionalidades foram integradas pelo uso do repositório online para o código fonte, utilizando o sistema Git. A modelagem e desenvolvimento das funcionalidades foram realizadas utilizando softwares abertos como StarUML e Eclipse IDE. A implementação das funcionalidades propostas teve como base outras ferramentas já utilizadas por pesquisadores para análise estatística e de correlação. Dentre estas ferramentas podemos destacar o pacote estatístico R e Apache Commons Math e Java Statistical Classes.

Para o desenvolvimento da nova funcionalidade da ferramenta foi feito um estudo bibliográfico dos métodos a serem desenvolvidos e análise da arquitetura atual da ferramenta base e para incorporar as novas funcionalidades. Em seguida, foi realizada a modelagem das novas funcionalidades. A validação dos resultados gerados

com a ferramenta foi realizada com comparativos usando ferramentas estatísticas de propósito geral como o ambiente R e Octave.

3. Desenvolvimento e Resultados

Foi definido que a funcionalidade seria implementada na tela de configuração de gráficos (Figura 1). Ao realizar uma avaliação de sistemas interativos, o usuário irá escolher se quer ou não utilizar os testes estatísticos. Os testes disponibilizados para essa funcionalidade foram: *Wilcoxon*, *Mann-Whitney U* e t-*de-Student.*

Na tela da Figura 1 é feita a escolha da configuração do gráfico, que é composto pelos sistemas/*runs* (os sistemas de recuperação da informação), consultas (uma busca que usuário submeteu ao sistema), iterações realizadas nas buscas e as medidas (utilizadas para mensurar a eficácia dos sistemas). Para criar um gráfico interativo com testes estatísticos é necessário marcar um *checkbox* localizado no canto inferior esquerdo e o gráfico deve atender algumas restrições: é necessário que o gráfico tenha apenas uma medida e no mínimo dois sistemas.

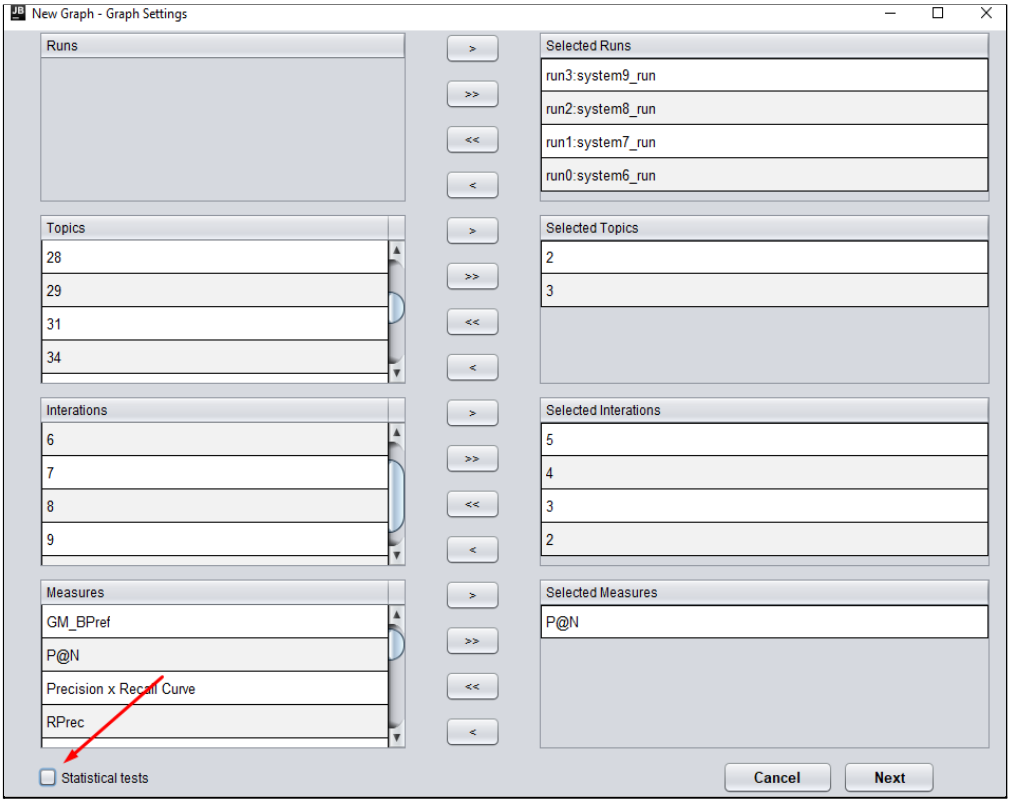

Figura 1. Tela de configuração de gráficos. No exemplo, os sistemas escolhidos foram *system9*, *system8*, *system7* e *system6*; as consultas (topics) foram a 2 e 3; as iterações: 2 a 5; e a medida *P@N*. Imagem retirada da ferramenta AnalyzIR.

Após a escolha da configuração do gráfico, o usuário é direcionado para a tela de configuração do teste estatístico, Figura 2, onde o usuário seleciona o tipo de teste,

dentre os disponíveis na ferramenta; o sistema (*run*), que servirá como *baseline* para ser comparado com os demais sistemas na análise estatística; e o nível de significância, indicado por *alpha*. Por fim, escolher a biblioteca a ser utilizada para o teste estatístico que também está disponível na ferramenta.

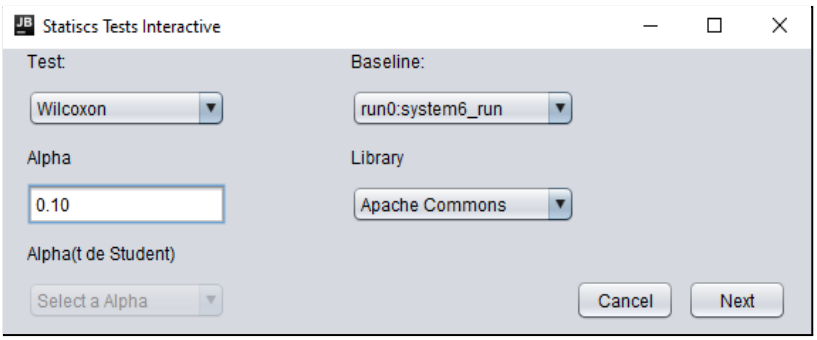

Figura 2. Tela de configuração do teste estatístico. No exemplo foi selecionado o teste de *Wilcoxon*. O *baseline* é o *system6*, a biblioteca é a *Apache Commons* e o nível de significância é de 0,10. Imagem retirada da ferramenta AnalyzIR.

Após definir as configurações de teste estatístico, o usuário terá que escolher o nome do gráfico e o seu título e depois a profundidade do *ranking* que será considerado na avaliação de cada sistema (*run*).

Assim, o gráfico será criado contendo o resultado da avaliação dos sistemas (*runs*) e o resultado do teste estatístico. Os dados que estão sendo apresentados são a média dos resultados das consultas, mas para o cálculo de significância estatística é utilizado o resultado da avaliação por consulta. O gráfico gerado pode ser de curva ou de barra. Nos gráficos de curva, para representar se houve ou não significância estatística foi utilizado os pontos. Caso não exista significância, a cor do ponto será igual ao da linha, mas caso exista, ela poderá ser azul para superioridade (quando o valor comparado é maior que o do *baseline*) ou vermelha para inferioridade (quando o valor comparado é menor que o do *baseline*), como representado na Figura 3.

O gráfico apresentado na Figura 3 inclui três sistemas: *system6*, *system7* e *system9*. Foram analisadas quatro iterações destes sistemas, sendo o *system6* definido como *baseline*. De acordo com o resultado da análise é possível observar que na segunda iteração houve significância para os sistemas *system7* e no *system9*, sendo que o *system9* é superior (indicado pelo ponto em azul) e o *system7* é inferior (indicado pelo ponto em vermelho). Também houve significância nas demais iterações dos dois sistemas citados anteriormente e todas apresentam inferioridade.

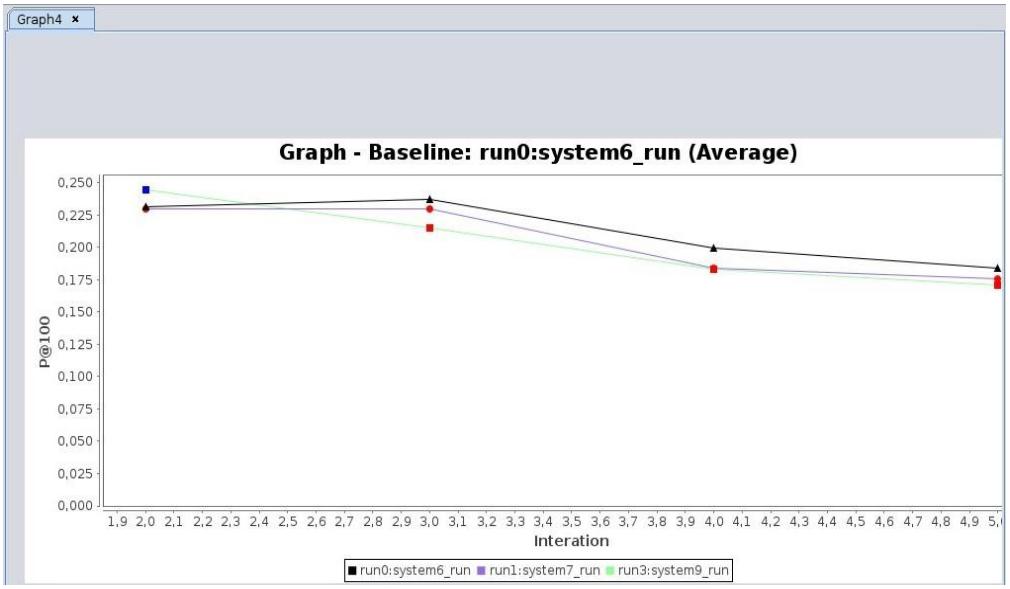

Figura 3: Exemplo de gráfico do tipo curva. Resultado da avaliação de eficácia e da análise de significância estatística de três sistemas ao longo de quatro iterações. Indicando significância nas quatro iterações dos sistemas comparados. Imagem retirada da ferramenta AnalyzIR.

Os gráficos de barra são utilizados quando o usuário só escolhe uma iteração. Caso exista significância, esta será indicada no número (localizado na parte superior da barra) que informa o valor encontrado na avaliação do sistema, seguindo o mesmo esquema de cor explicado anteriormente, mudando apenas no cenário onde não houve significância, a cor representada será a preta. Além disso, na legenda que informa que cor representa cada run haverá conter um "*" no(s) sistema(s) com significância estatística. Um exemplo de gráfico de barra está representado na Figura 4.

O gráfico da Figura 4 apresenta 4 sistemas. Foi analisada apenas uma iteração dos sistemas, e o *baseline* escolhido foi o *system1*. O resultado indica que houve significância para os sistemas *system2* e *system4*, como é possível observar na legenda, onde esses dois sistemas apresentam um "*" ao lado do nome, e pela legenda do valor encontrado na avaliação de eficácia que está na cor vermelha, indicando inferioridade.

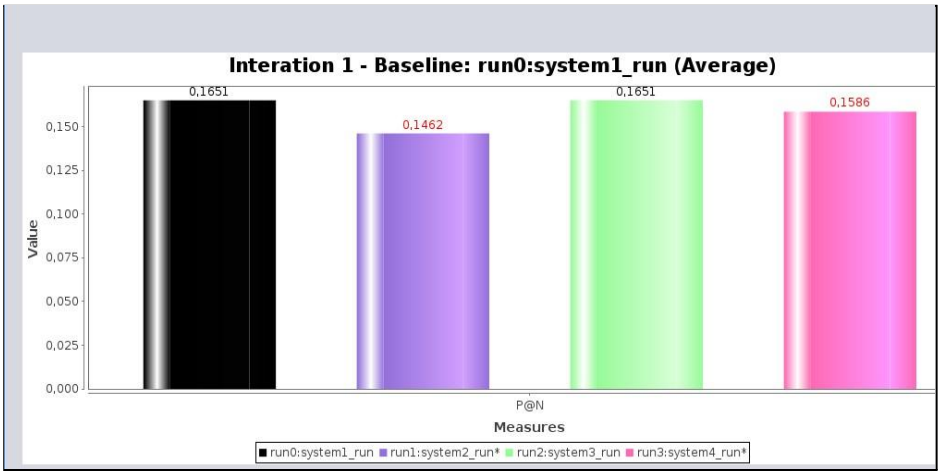

Figura 4: Exemplo de gráfico do tipo barra. Resultado da avaliação de eficácia e da análise de significância estatística de quatro sistemas e apenas uma iteração. Indicando significância para dois dos três sistemas comparados. Imagem retirada da ferramenta AnalyzIR.

É possível exportar o gráfico nos formatos: jpg, png, xls e csv. O formato xls contém os dados do resultado da análise do sistema (*run*) e do teste estatístico.

4. Considerações Finais

Devido a grande parte da bibliografia dos assuntos abordados no projeto e da documentação da ferramenta AnalyzIR ser em inglês, e não ter um completo domínio dessa língua, no início do desenvolvimento do projeto houve um pouco de dificuldade na compreensão de alguns aspectos, mas esse obstáculo auxiliou na melhora entendimento da língua, e com isso ajudando em trabalhos posteriores. Após o projeto finalizado, é possível que o usuário aplique o teste em sistemas interativos e tenha um retorno visual tanto do resultado da análise do sistema como do teste estatístico. Além disso, é possível exportar os dados de avaliação do sistema e os da análise estatística para serem estudados posteriormente, assim, se tornando uma ferramenta mais completa para o uso dos pesquisadores. Uma possível melhoria nessa nova funcionalidade seria encontrar uma solução para apresentar os resultados divididos por tópicos, pois atualmente para manter a legibilidade do gráfico é apresentado o resultado da média dos tópicos, assim trazendo ainda mais material de estudo para os pesquisadores.

5. Referências Bibliográficas

Calumby, R., Gonçalves, M., Torres, R. (2016) "On Interactive Learning-to-Rank for IR: Overview, Recent Advances, Challenges, and Directions. Neurocomputing". Amsterdam

Calumby, R. (2010) Recuperação Multimodal de Imagens Com Realimentação de Relevância Baseada em Programação Genética. Instituto de Computação, Unicamp.

Torres, R., Falcão, A. (2006) Content-based image retrieval: Theory and applications. Revista de Informática Teórica e Aplicada, 13(2):161–185.

Penatti, O. (2009) Estudo Comparativo de Descritores para Recuperação de Imagens por Conteúdo na Web. Instituto de Computação, Unicamp.