Um estudo sobre o uso do modelo probabilístico no desenvolvimento de uma máquina de busca

Stephany Castro da SILVA[1];

stephanycastro.es@gmail.com;

Aurélio Andrade de Menezes JÚNIOR[1].

aurelio.menezesjr@gmail.com

[1]Universidade Federal do Amazonas (UFAM), Instituto de Ciências Exatas e Tecnologia (ICET), Itacoatiara, AM, Brasil.

UM ESTUDO SOBRE O USO DO MODELO PROBABILÍSTICO NO DESENVOLVIMENTO DE UMA MÁQUINA DE BUSCA

Resumo

Este trabalho propôs a criação de uma base de dados que contivesse de forma digitalizada as devidas Resoluções existentes e propôs também a realização de um estudo sobre o uso do Modelo BM25 no desenvolvimento de uma Máquina de Busca voltada para resoluções instituicionais utilizadas pela Universidade Federal do Amazonas, como um meio de minimizar o problema de acesso e recuperação de tais instrumentos normativos.

Palavras-chave: Modelo BM25; Resoluções; Máquina de Busca**.**

Abstract

This paper proposes the creation of a database that contained digitized so appropriate existing resolutions and also proposed to conduct a study on using the BM25 model in developing a focused search machine for institucional resolutions used by the Federal University of Amazonas as a means of minimizing the problem of access and retrieval of such normative tools.

Keywords: Model BM25; Resolutions; Search engine.

Introdução

Com o advento da Web e de sua grande utilização, a quantidade de banco de dados existentes em instituições, tais como, escolas, empresas e associações, cresceu rapidamente. Isso aumentou a quantidade de dados existentes e de informações trocadas de forma eletrônica. De acordo com (Júnior, 2012), a Web tem revolucionado tanto o acesso a informações pessoais quanto o gerenciamento do conhecimento em instituições, assim como bancos de dados tem sido usado com grande frequência no nosso dia a dia nas empresas para armazenamento e controle de informações.

Entretanto, se hoje temos a facilidade de acessar grandes quantidades de informações, por outro lado, também temos um problema causado por isso: Como encontrar uma determinada informação dentro de grandes quantidades de Informações. Para resolver esse problema utiliza-se a Recuperação da Informação - RI.

RI é uma área responsável pela representação, armazenamento, organização e acesso a itens de informação (Garcia, 2002). Além disso, é uma ciência de pesquisa relacionada à busca por informações que podem estar contidas em determinados documentos, busca pelos documentos desejados pelo usuário e busca em banco de dados, este último podendo ser relacionais e isolados, ou até mesmo interligado a Web. Esses documentos podem conter qualquer tipo de mídia, tais como, texto, imagem e som. Na maioria das vezes, são compostos de texto em linguagem natural ou de informação textual associada a outros tipos de dados.

Dentro desse contexto, há as Máquinas de busca que existem para facilitar o acesso a informações relevantes de acordo com o desejo do usuário. Máquinas de busca são projetadas para operarem ambientes onde a quantidade de conteúdo disponível

supera a capacidade do usuário de acessá-lo de forma eficiente durante a pesquisa (Júnior, 2012). Esses sistemas são implementados a partir de Modelos em Recuperação de Informação que são o núcleo de qualquer sistema de RI. Tais modelos são utilizados para representar características semânticas dos elementos envolvidos nos sistemas, sendo os três tipos de modelos para Recuperação de Informação que possuem o seu uso mais difundido: o modelo booleano, o modelo vetorial e o modelo probabilístico.

O modelo booleano é composto por conjuntos de documentos e operações clássicas da teoria de conjuntos. O modelo vetorial é composto por documentos e consultas representados como vetores em um espaço n dimensional e operações de álgebra linear aplicáveis aos vetores (Oliveira, 2010). O modelo probabilístico define as representações para documentos e consultas baseado na teoria das probabilidades. A base desse modelo está no princípio da ordenação probabilística: onde dada uma consulta e um documento, tenta-se estimar a probabilidade do usuário considerar o documento relevante à consulta (Oliveira, 2010).

Baseado em tais modelos clássicos, outros modelos, ou funções específicas foram sendo criados, tais como, a função BM25 que foi desenvolvido baseado em modelos probabilísticos. Tal função de classificação é utilizada para atribuir relevância aos documentos existentes em uma Máquina de Busca, nele a busca de uma informação é baseada na probabilidade de um documento ser relevante para a consulta (Martins, 2009).

A função BM25 é baseada no conceito de utilização de uma “bolsa” de palavras para classificar um conjunto de documentos de acordo com a sua relevância para uma determinada consulta, independentemente da inter-relação entre os termos da consulta em um documento. O desempenho de um sistema de RI é avaliado de acordo com a sua

capacidade em recuperar o maior número de itens relevantes, ao mesmo tempo em que filtra ao máximo os itens irrelevantes (Oliveira, 2010).

Sendo o objetivo de uma máquina de busca facilitar o acesso a informação relevante, conhecer a sua eficiência é importante não só para os pesquisadores de RI, mas também para quem usa estes sistemas. Pesquisadores e usuários precisam ter maneiras efetivas para saber quão bons são os sistemas para uma dada tarefa e como estes podem ser melhorados.

Algumas medidas mais comuns para avaliar o desempenho de um sistema computacional são tempo e espaço. Se o tempo de resposta do sistema for pequeno e o tamanho do espaço de memória utilizado também, o sistema pode ser considerado excelente. Entretanto, quando se trata de sistemas cujo objetivo é recuperar informações outras medidas devem ser utilizadas.

Sendo assim, para determinar se um sistema de busca é bom ou não, utiliza-se Métodos de Avaliação específicos, que verificam um sistema de RI através da comparação das respostas geradas por este e o conjunto ideal de respostas esperadas. Isto fornece uma estimativa da qualidade do algoritmo de recuperação de informação avaliado. As métricas usuais para a avaliação do resultado de um sistema de recuperação de informação são: revocação e precisão. Para que estas medidas sejam relevantes, é necessário conhecer bem o conteúdo dos documentos da coleção (Oliveira, 2010). Revocação mede a proporção de documentos relevantes que formam retornados como resultado a uma consulta do usuário e a Precisão mede quantos documentos relevantes foram recuperados

Material e Métodos

O projeto foi elaborado de acordo com as etapas previstas no cronograma. As atividades realizadas foram referentes à:

Realizar levantamento bibliográfico de trabalhos relacionados a métodos de avaliação. Foram estudadas e analisadas as técnicas e aplicações de um método de avaliação em uma Máquina de Busca. Assim, foi possível conhecer a fundo esses determinados tipos de métodos, verificar a aplicação de um método de avaliação de desempenho na Máquina de Busca, para aplicar tais métodos na prática para a validação da Máquina de Busca criada.

Analisar o comportamento do Modelo BM25. Foi avaliada o funcionamento do Modelo BM25, por meio de estudos dos conceitos referentes a esse tipo de modelo e verificado a sua aplicação no desenvolvimento de uma Máquinas de Busca.

Realizar levantamento de resoluções da UFAM que possam ser digitalizadas para a criação de uma base de dados e analisar os conceitos referentes à implementação de uma Máquina de Busca. Foram levantadas resoluções que podiam ser digitalizadas para a criação de uma base de dados. Após isso, foram estudados os conceitos sobre Máquina de Busca e sobre o funcionamento de tais sistemas de RI, para por em prática e implementar o Modelo BM25 para aplicá-lo em uma Máquina de Busca.

Estudar e analisar as técnicas e aplicações de um método de avaliação em uma Máquina de Busca. Foram realizados estudos referentes a esses determinados tipos de métodos. Além disso, foi verificado métodos de avaliação de desempenho para serem aplicadas na validação da Máquina de Busca desenvolvida.

Resultados

Durante os primeiros meses de execução do projeto, foi realizado levantamento bibliográfico de trabalhos e artigos relacionados a área do estudo. Os assuntos levantados foram referentes à Recuperação da Informação, dentro da qual, foram realizados estudos abordando o Modelo Probabilístico (modelo base para a construção do Modelo BM25), a qual o mesmo, foi utilizado durante o decorrer do trabalho, mais precisamente para a avaliação do seu uso no desenvolvimento de uma Máquina de Busca.

Sendo um dos objetivos do projeto a implementação do Modelo BM25 como uma forma de facilitar e agilizar o processo de consulta de um conjunto de resoluções usadas pela Universidade Federal do Amazonas – UFAM, fez-se necessário a construção de uma base de dados contendo resoluções, obtidas a partir da página oficial da Instituição. Os textos obtidos foram posteriormente todos convertidos para o formato .txt, dado a maior facilidade em aplicar as buscas em documentos com este formato.

Com a compreensão dos principais assuntos relacionados ao Modelo BM25, foi possível criar uma versão do mesmo, capaz de executar os principais passos do processo de recuperação da informação que são: consulta, indexação e pesquisa. A consulta consiste na especificação de um conjunto de termos que representam a necessidade de informação do usuário, a indexação envolve a criação de estruturas de dados associadas aos documentos de uma coleção, onde se encontra a possível informação que foi requerida pelo usuário. A pesquisa envolve o processo de recuperar os documentos de acordo com a consulta do usuário. A ordenação dos documentos recuperados dos mais relevantes para o menos relevantes, é feito usando Modelos como o BM25.

Como pode ser observado na **Figura 1** e **Figura 2**, a primeira etapa de funcionamento da estrutura da máquina de busca, consiste na Busca informada pelo usuário, ou seja, é a entrada dos termos para a pesquisa das resoluções na base onde estão as resoluções levantadas em etapas anteriores do projeto. A segunda parte da busca é feita aplicando graus (valores) de relevância na coleção como pode ser observado na **Figura 3**, que permitem posteriormente o Modelo BM25, classificar cada documento relevante para a consulta e por fim realizar a ordenação de acordo com a relevância. Assim, quanto maior o grau de relevância do documento, mais importante ele é para o usuário.

As métricas utilizadas para avaliação da eficiência da Máquina de Busca desenvolvida por este projeto, foram a de precisão e revocação. Tais métodos foram selecionados a partir dos estudos realizados sobre o assunto, onde identificou-se que ambas são métricas comumente utilizadas dentro da área de Recuperação da Informação, como forma de avaliar rankings de busca. O processo de avaliação da estratégia utilizado para essa avaliação, seguiu os seguintes passos:

1. Criação da lista de consultas com as palavras-chave: a criação da lista iniciou-se com a escolha de 11 palavras – chave, as mesmas foram selecionados após a realização de análises na base de dados onde se identificou as palavras que mais se repetiam durante o decorrer do texto.

2. Identificação dos documentos mais relevantes: foram identificados os documentos relevantes e foi atribuído um peso de 1 a 8, onde 8 indica o documento mais relevante e 1 indica o documento menos relevante a uma dada consulta.

3. Execução da busca no sistema de RI desenvolvido: utilizando as palavras-chave da lista de consulta pré-definidas, o algoritmo foi executado e obteve-se a lista de resultado por ele gerado.

4. Análise dos resultados: o resultado da busca do Modelo BM25 foi comparado com os documentos relevantes da lista de consulta.

Discussões

Este trabalho propôs a criação de uma base de dados que contivesse de forma digitalizada as devidas Resoluções existentes e propôs também a realização de um estudo sobre o uso do Modelo BM25 no desenvolvimento de uma Máquina de Busca voltada para resoluções instituicionais utilizadas pela Universidade Federal do Amazonas, como um meio de minimizar o problema de acesso e recuperação de tais instrumentos normativos.

A aplicação das métricas precisão e revocação como meio de avaliar a máquina de busca desenvolvida utilizando o Modelo BM25, observou-se que a média de precisão do algoritmo de 0,710, com a revocação de 0,043. Esses resultados permitiram verificar que quanto maior a precisão do algoritmo de busca, menor será a sua revocação, além disso, também foi possível perceber que o Modelo BM25 concentra seus resultados mais relevantes próximos do topo.

Como resultado, pudemos identificar que a abordagem aplicada neste trabalho gerou precisão de 71% e revocação 4,3%, além de apresentar tempo médio de resposta de 58 milissegundos. Um fator que pode ter influenciado no resultado obtido é que como a maioria das resoluções disponíveis para acesso via web estava no formato de

imagem, a conversão dos mesmos para texto pode ter causado uma certa perda de dados importantes para a busca.

Entretanto, os resultados do experimento foram satisfatórios e serviram para demonstrar a relevância deste trabalho e além disso, permitiu a obtenção de conhecimentos importantes na área da Recuperação da Informação, principalmente a assuntos referentes ao Modelo BM25 e aos Métodos de Avaliação.

Agradecimentos

Os autores agradecem a FAPEAM (Fundação de Amparo a Pesquisa do Amazonas) pelo apoio financeiro, e a todos aqueles que contribuíram e que continuarão contribuindo de alguma forma para o sucesso deste trabalho.

Bibliografia Citada

Ferreira, Luciene Costa. 2009. *Repositório Temático Digital dos Instrumentos Normativos da UFBE: Acesso e Recuperação Da Informação Através da Plataforma DSPACE.* Universidade Federal da Paraíba.

Garcia, Tiago Freire. 2002. *Proposta de uma Máquina de Busca Eficiente para Documetos na Web Usando Lógica Fuzzy.* Monografia. Universidade Federal de Lavras. 3p.

Júnior, Aurélio Andrade de Menezes. 2012. *Um Método para Busca de Competências a Partir de currículos Lattes*. Dissertação de Mestrado. Universidade Federal do Amazonas. Instituto de Computação. Programa de Pós-Graduação em Informática. 12p.

Oliveira, Renan Rodrigues. 2010. *Recuperação Contextualizada de Documentos Integrados pelo Protocolo OAI-PMH*. Dissertação de Mestrado. Universidade Federal de Goiás. Instituto de Informática. Programa de Pós-Graduação do Instituto de Informática da Universidade Federal de Goiás. 44-45p.

Figuras

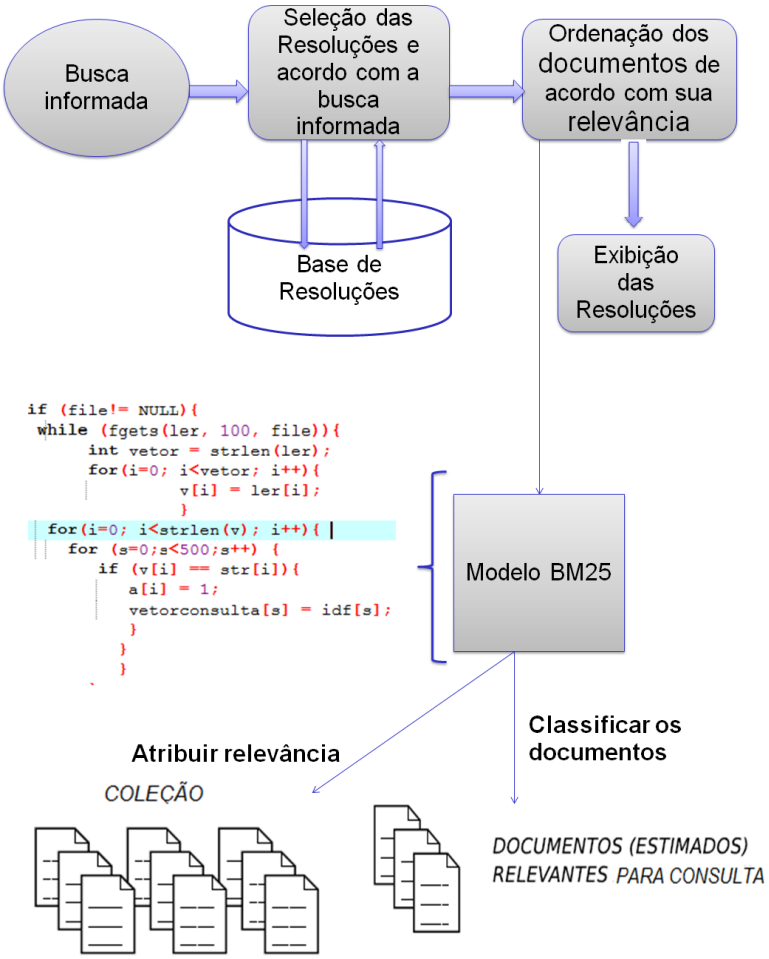

Figura 1: Estrutura do funcionamento da Máquina de busca, usando o Modelo BM25.

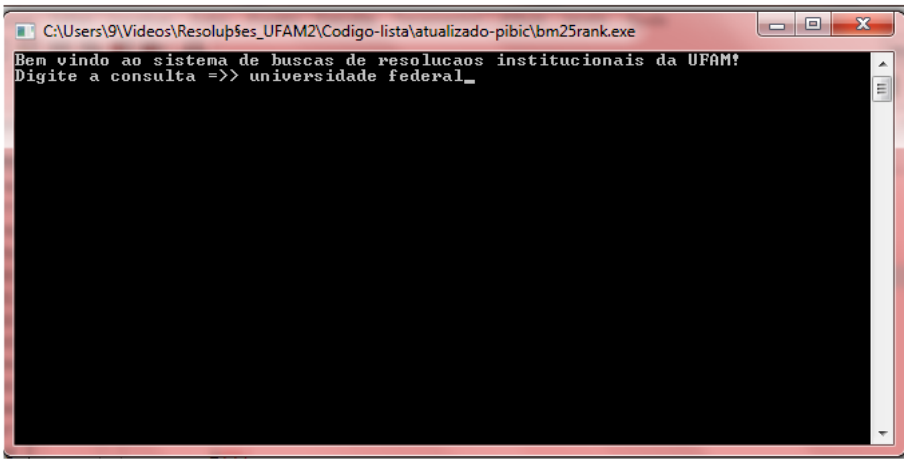

Figura 2: Consulta usando o Modelo BM25.

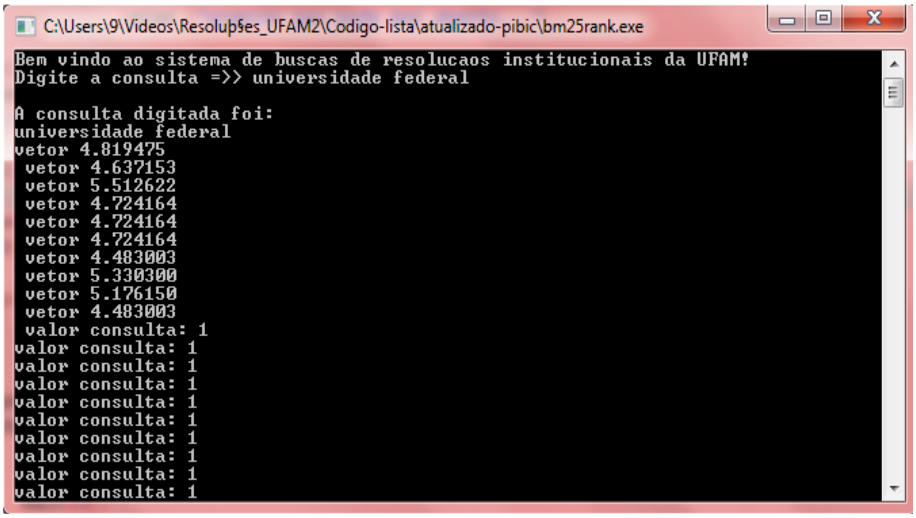

Figura 3: Aplicação dos graus (valores) de relevância na coleção de resoluções usando o Modelo BM25.

Tabelas

Nº	Descrição	Ago 2014	Set	Out	Nov	Dez	Jan 2015	Fev	Mar	Abr	Mai	Jun	Jul
1	Converter as resoluções levantadas de imagem para formato txt	X	X										
2	Implementar a máquina de busca utilizando o modelo bm25	X	X	X	X	X							
3	Aplicação de métodos de avaliação de desempenho para a verificação da funcionalidade da máquina de busca desenvolvida					X	X	X	X				
4	Elaboração do Resumo e Relatório Final									X	X	X	X
5	Preparação da Apresentação Final para o Congresso											X	X