Proposta e desenvolvimento de uma estratégia de recuperação de informação em um banco de questões do opCoders Judge

Autor: Felipe Braz Marques

21 de setembro de 2024

1 Contextualização e Definição do Problema de Pesquisa

Com a evolução da tecnologia, os olhares do mundo voltaram para a programação. E nos dias atuais, a disciplina de Programação de Computadores é parte obrigatória da grade curricular de muitos cursos da área de ciências exatas, com o objetivo de desenvolver o raciocínio lógico de que o aluno necessitará durante o curso. Estas disciplinas, que envolvem o primeiro contato com a programação, são consideradas muito difíceis; com isso, surgiu a necessidade da criação de um corretor automático que auxilie tanto os professores quanto os alunos, nascendo o opCoders Judge que de acordo com (PATROCINIO, 2023)foi criado em duas etapas por graduandos de Ciência da Computação na Universidade Federal de Ouro Preto: na primeira etapa, uma versão offline para a correção das tarefas dos alunos na máquina local do professor foi desenvolvida; e na segunda etapa, foi implementada uma versão Web, onde o aluno visualiza a questão pela página Web e submete pela mesma os códigos fontes.

Como opCoders Judge é uma ferramenta nova, existem inúmeras possibilidades de aprimorá-lá. Uma delas seria a criação de um banco de dados contendo questões mais específicas para cada tipo de curso, ocasionando com que o aluno se sentisse mais motivado em aprender, visto que as questões que ele resolveria estariam mais perto da sua área de interesse. Exemplificando, um aluno do curso de Engenharia Civil poderia resolver exercícios relacionados a cálculos estruturais de pontes ou edifícios, um aluno de Engenharia Metalúrgica a respeito do tratamento de materiais metálicos em processos de uma siderurgia, e um aluno de Engenharia de Minas a respeito de propriedades físicas de uma determinada gema ou sobre processos de extração de um determinado mineral.

Considerando um banco de questões, para uma melhor resposta ao se buscar por alguma questão relacionada a algum assunto ou tópico específico no mesmo, ou seja, uma lista de questões relevantes ao interesse de um aluno em particular, é interessante a aplicação de conhecimentos envolvendo uma área da Ciência da Computação denominada Recuperação de Informação (RI). De acordo com (ALVAREZ; GONÇALVES, 2017), a RI visa garantir a qualidade dos resultados retornados, de forma rápida, mediante quaisquer consultas. Complementando, segundo (WIVES, 1997), RI refere-se ao ato do usuário especificar e descrever a informação de que ele precisa, juntamente com os métodos utilizados para recuperar essas informações. Já (BAEZA-YATES; RIBEIRO-NETO, 1999) define a RI como sendo a área da Ciência da Computação que estuda a recuperação de informação (não dados) de uma coleção de documentos, visando satisfazer as necessidades do usuário normalmente expressas em linguagem natural; para tanto, a RI lida com a representação, o armazenamento, a organização e o acesso a itens de informação.

A representação e a organização dos itens de informação provêem um acesso fácil à é realmente relevante para os usuários, visto que um usuário pode não possuir uma informação

de interesse do usuário (MEIRA et al., 2002). Entretanto, não é uma tarefa fácil determinar o que é realmente relevante para os usuários, visto que um usuário pode não possuir uma descrição bem detalhada sobre seu objeto de consulta. Tal descrição, geralmente, é composta por termos chaves, que sintetizam a informação desejada. Os termos chaves referem-se aos termos de indexação que permitem a recuperação de documentos, por exemplo, por máquinas de busca; cada documento é representado por um conjunto desses termos, em que as semânticas dos mesmos especificam o conteúdo do documento.

De acordo com (ALVAREZ; GONÇALVES, 2017), os algoritmos de classificação são os responsáveis por decidir, mediante uma consulta, quais documentos são relevantes ou não à mesma, devido ao fato desses algoritmos tentarem estabelecer uma ordem para os documentos recuperados, sendo que os documentos que aparecem no topo da lista são considerados os mais relevantes. Essa classificação dá-se de acordo com um conjunto de premissas definido para o modelo de RI que esteja sendo utilizado pelo algoritmo de classificação. Dessa forma, o modelo de RI adotado determina a suposição do que é ou não relevante, mediante uma consulta.

Com o propósito de proporcionar respostas rápidas e de qualidade, alguns modelos de RI chamados clássicos foram propostos: o Booleano, o Vetorial e o Probabilístico. O modelo Booleano restringe-se apenas a comparar, por meio de operadores lógicos, os termos dos documentos de uma determinada coleção com os termos fornecidos pelo usuário em uma consulta. Já o Vetorial e o Probabilístico utilizam-se de estratégias mais avançadas para ordenar os documentos retornados de uma coleção em relação à relevância dos mesmos para com a consulta fornecida. De uma forma geral, sabe-se que o modelo Booleano é o mais limitado dos três e que os modelos Vetorial e Probabilístico apresentam bons resultados na RI, possuindo comportamentos bem próximos.

Visando a melhoria da qualidade dos resultados gerados pela aplicação dos modelos eles, tem-se o Extended Boolean, o Generalized Vector e o Belief Network. Os modelos clássicos de RI, foram definidos, a partir dos mesmos, modelos estendidos de RI (HIEMSTRA; VRIES, 2000); dentre eles, tem-se o Extended Boolean, o Generalized Vector e o Belief Network. Os modelos estendidos proporcionam uma maior qualidade na recuperação de informação, mas podem ser inviabilizados por diferentes motivos. Segundo(KRAAIJ, 2004) o Extended Boolean não é usual porque a montagem das consultas é complexa, embora a utilização de interfaces de apoio juntamente com consultas curtas poderia torná-lo um modelo atrativo. O Generalized Vector, por sua complexidade computacional, só é aplicável para consultas curtas em uma pequena quantidade de documentos. Já o Belief Network tem seu comportamento ditado pela estratégia de classificação escolhida. Dessa forma, acredita-se que cada um dos modelos estendidos seja melhor aplicável a uma situação específica.

Portanto, para tentar amenizar as taxas de reprovações e as desistências em disciplinas, o opCoders Judge pode se tornar ainda mais amigável a partir da utilização de modelos de RI para localizar questões relevantes de um banco de questões em relação ao interesse do aluno. Diferente de alguns sistemas de correção automática disponíveis na web como Beecrowd, que é um portal de acesso gratuito que disponibiliza desafios de programação e corrige respostas instantaneamente como sugere (CEDRAZ, 2023), no opCoders Judge, o aluno de cada curso faz questões que abordam assuntos da sua respectiva área, fazendo despertar o interesse na programação. E, para alcançar tal objetivo, no intuito de se produzir conhecimento, é necessário aplicar conhecimentos de algumas áreas da Ciência da Computação, como RI, no ambiente já existente.

2 Objetivos

Este projeto de pesquisa possui, como objetivo principal, a proposta, o desenvolvimento e a validação de uma estratégia para localizar com eficácia e eficiência, dentre as questões presentes no banco de dados do opCoders Judge, aquelas que sejam relevantes a um determinado assunto ou tópico específico. Para o cálculo de similaridade entre o que se deseja e as questões do banco, serão considerados, em princípio, os modelos clássicos e estendidos de RI; ademais, um dicionário de termos relativos ao contexto da coleção ou um tesauro podem ser utilizados para melhorar a determinação da relevância. Para validar a ferramenta proposta, experimentos serão realizados, envolvendo distintas consultas e o banco de questões.

Os objetivos específicos a serem atingidos são:

- implementação e validação prática de distintos modelos de RI;
- estudo comparativo da eficácia e da eficiência dos modelos de RI implementados, com base nos resultados experimentais obtidos;
- levantamento de dados estatísticos quanto aos termos presentes no banco de questões do opCoders Judge;
- consolidação da linha de pesquisa Tratamento e Recuperação da Informação do Departamento de Computação da Universidade Federal de Ouro Preto.

3 Hipótese de Trabalho

Esse projeto de pesquisa propõe melhorias para o opCoders Judge, a partir da implementação de uma estratégia de RI que, a partir de um assunto ou tópico específico, busca em um banco de questões aquelas que tenham certo grau de similaridade. Desta forma, a ferramenta é útil para várias situações, a saber: (a) buscar questões de uma disciplina que estejam relacionadas ao curso de determinados alunos, no intuito de gerar provas específicas para os mesmos; (b) buscar questões relacionadas a questões acertadas ou erradas por um determinado aluno, no intuito de aprimorar e enriquecer seu conhecimento; e (c) buscar questões similares que possam contribuir em um processo de aprendizagem adaptativa de alunos.

Quanto ao item (a) apresentado, como situação real, atualmente as provas das disciplinas de Programação de Computadores da Universidade Federal de Ouro Preto são oferecidas no papel e de forma igualitária para todos os alunos matriculados nas mesmas, independente de seus cursos. A partir da estratégia de RI proposta neste trabalho, tornase possível buscar questões no banco de dados relacionadas aos cursos dos alunos no intuito de produzir provas mais direcionadas aos seus próprios cursos, estimulando os mesmos na realização da atividade. Assim, a estratégia proposta auxiliará diretamente na resolução de um problema real na UFOP, além de ajudar os alunos a se sentirem mais confortáveis com a programação.

Em resumo, as principais e possíveis contribuições deste projeto são:

- confecção de uma estratégia para cálculo de similaridade entre um assunto e questões no banco de dados do opCoders Judge, podendo ser utilizada em situações reais;
- realização de um conjunto extenso de experimentos, envolvendo distintas consultas e o banco de questões, no intuito de validar a estratégia proposta;
- geração de uma nova e mais aprimorada versão do opCoders Judge.

4 Procedimento Metodológico

De uma forma geral, este projeto prevê as seguintes atividades:

- revisão de literatura sobre modelos de RI e sobre a tecnologia para implementação da estratégia proposta;
- implementação de modelos clássicos de RI;
- implementação de modelos estendidos de RI;
- incorporação de um dicionário de termos ou de um tesauro nos modelos de RI implementados;
- realização de experimentos de validação dos modelos implementados;
- definição da estratégia de RI a ser utilizada no opCoders Judge a partir da análise dos resultados obtidos nos experimentos;
- incorporação da nova versão do opCoders Judge em aplicações reais de grande vulto;
- utilização da nova versão do opCoders Judge em aplicações reais de grande vulto.

Referências

ALVAREZ, G. M.; GONÇALVES, A. L. Qualidade da Informação e Recuperação de Informação: uma revisão da literatura. **Revista Tecnologia da Informação e Comunicação: Teoria e Prática (UNISUL)**, v. 1, n. 1, 2017.

BAEZA-YATES, R. A.; RIBEIRO-NETO, B. A. **Modern Information Retrieval**. New York, NY: ACM Press/Addison-Wesley, 1999.

CEDRAZ, V. F. **Uma avaliação de usabilidade do corretor de exercícios de introdução à programação OpCoders Judge**. 2023. Universidade Federal de Ouro Preto, Ouro Preto. Monografia (Graduação em Ciência da Computação). Disponível em: <https://www.monografias.ufop.br/handle/35400000/5507>.

HIEMSTRA, D.; VRIES, A. P. **Relating the new language models of information retrieval to the traditional retrieval models**. Enschede, 2000. Disponível em: <http://wwwhome.cs.utwente.nl/~hiemstra/papers/tr-ctit-00-09.pdf>.

KRAAIJ, W. **Variations on Language Modeling for Information Retrieval**. 2004. Tese (Doutorado) – Taaluitgeverij Neslia Paniculata / CTIT Ph.D. thesis series, Enschede. Disponível em: <http://dis.tpd.tno.nl/mmt/pubs/wkthesis.pdf>.

MEIRA, W. et al. Set-Based Model: A New Approach for Information Retrieval. In: PROCEEDINGS of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval. Tampere, Finland: [s.n.], 2002.

PATROCINIO, J. A. do. **OpCoders Judge: Uma versão online para o corretor automático de exercícios de programação do projeto opCoders**. 2023. Universidade Federal de Ouro Preto, Ouro Preto. Monografia (Graduação em Ciências da Computação). Disponível em: <https://www.monografias.ufop.br/handle/35400000/5360>.

WIVES, L. K. **Um Estudo Sobre Técnicas de Recuperação de Informações Com Ênfase em Informações Textuais**. 1997. Tese (Doutorado) – Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre. Disponível em: <http://www.inf.ufrgs.br/~wives/publicacoes/TI.pdf>.