

On6tão – Coleta e integração de dados de egressos

CHRISTIAN FELIPE ARAUJO TEIXEIRA (Autor), ANDERSON ALMEIDA FERREIRA (DECOM) (Orientador), FABIO JOSE DE MEDONCA JUNIOR (Co-Autor)

Instituição de Ensino - Universidade Federal de Ouro Preto

Palavras Chaves:

resolução de entidades; on6tão; entity linking; tratamento da informação

Resumo:

Um curso de graduação precisa preparar seus alunos para os possíveis desafios e barreiras que podem surgir após a sua conclusão. Uma das maneiras de identificar quais são essas demandas é por meio da gestão de informações dos egressos, ou seja, coletar e analisar dados profissionais e acadêmicos de egressos. Este trabalho é parte de um projeto maior, chamado On6tão, que visa analisar dados de egressos usando informações disponíveis publicamente em redes sociais online. Centrado no módulo Gerência de Dados do On6tão, que é composto por três componentes chamados de Coleta, que obtém páginas web com informações sobre os egressos, Extração, que extrai os dados dessas páginas, e Integração, que integra e padroniza esses dados, este projeto implementa esses dois últimos componentes. Utilizando duas fontes de dados, LinkedIn e Facebook, a aplicação desenvolvida recebe como entrada uma coleção de páginas do componente Coleta e produz, como resultado, um conjunto de registros contendo os dados integrados dos egressos. Para a extrair os dados das páginas, utilizou-se padrões definidos por meio de expressões regulares, visto que há regularidade na estrutura das páginas de cada fonte, e implementou-se um algoritmo para extrair esses dados identificados por meio desses padrões. Para o componente Integração, foram avaliadas métricas de similaridade e utilizadas bases de conhecimento externas, como a Wikipédia e o Google Maps, para integrar as informações obtidas pertencentes aos egressos. Como resultados obtidos, a extração obteve quase a totalidade dos atributos dos egressos da fonte LinkedIn, enquanto do Facebook obteve, em média, 54% dos atributos. A resolução de entidades na integração teve um acerto médio em torno de 72%. Como trabalho futuro, pretende-se realizar experimentos com dados de egressos de outros cursos e instituições e investigar formas de aprimorar a resolução de entidades. Este trabalho teve o apoio da FAPEMIG via edital PROBIC e da UFOP.

Publicado em:

- Evento: Encontro de Saberes 2016
- Área: CIÊNCIAS EXATAS E DA TERRA
- Subárea: CIÊNCIA DA COMPUTAÇÃO