Uma estratégia para recomendação de especialistas a partir de dados abertos disponíveis na Plataforma Lattes

Sérgio José de Sousa
Mestrando em Modelagem Matemática e Computacional pelo Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) – Belo Horizonte, MG, Brasil. Graduação em Sistemas de Informação pela Faculdade Pitágoras de Divinópolis (FAP) - Brasil.
http://lattes.cnpq.br/1639967799540564
E-mail: sergio7sjs@gmail.com

Thiago Magela Rodrigues Dias
Doutor em Modelagem Matemática e Computacional pelo Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) – Belo Horizonte, MG - Brasil. Professor do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) – Belo Horizonte, MG, Brasil.
http://lattes.cnpq.br/4687858846001290
E-mail: thiagomagela@gmail.com

Adilson Luiz Pinto
Pós-Doutorado pelo Institut de Recherche en Sciences de l´Information et de la Communication (IRSIC) - França. Doutor em Documentación pela Universidad Carlos III de Madrid (UC3M) - Espanha. Professor da Universidade Federal de Santa Catarina,(UFSC) - Florianópolis, SC - Brasil.
http://lattes.cnpq.br/4767432940301118
E-mail: adilson@cin.ufsc.br

Data de submissão: 13/09/2019. Data de aprovação no ConfOA: 11/06/2019. Data de publicação:

RESUMO

Com o crescente volume de dados é um desafio cada vez maior encontrar informações que se desejam. Neste contexto, agentes de recomendação e técnicas de ranqueamento são considerados boas soluções para o problema da sobrecarga de informação.Os xmls dos currículos são extraídos. Logo em seguida os dados são selecionados, extraindo títulos de artigos de periódicos, trabalhos de anais, congressos e projetos de pesquisa, após o que indexamos todos os termos, totalizando e atribuindo um valor inteiro sequencial para cada palavra. Isto totaliza mais de 307 mil currículos e mais de 1 milhão de termos encontrados. Retiramos então palavras menos significativas, como stopwords, aplicamos radicalização (stemming) e extraímos o radical da palavra. Já ao gruparmos os termos iguais, totalizamos 68.417 radicais. Agora, cada especialista é representado por um vetor de tamanho 68.417, e cada valor é calculado por meio de TF-IDF, que leva em conta a frequência do termo e o quão comum ele é, valorizando mais o termo menos usado. Por fim, esse dado é utilizado para treinar um Autoencoder e, decorrência de uma entrada, ele é reduzido até 25 dimensões, antes do que ele é ampliado novamente, com o objetivo de minimizar o erro de reconstrução. De posse desses novos dados, podemos utilizar o mesmo Encoder para transformar consultas nesse espaço de 25 dimensões e, assim, computar a distância entre os especialistas e a consulta.

Palavras-chave: Recomendação de Especialista. Plataforma Lattes. Redes Neurais. Deep Learning.

A strategy for the recommendation of experts from open data available on the Lattes platform

ABSTRACT

With the increasing volume of data, it is increasingly difficult to find the information you want. In this context, recommendation agents and classification techniques are considered good solutions to the problem of information overload. The xmls of the curricula are extracted. Shortly after, the data is selected, extracting titles of articles from journals, anal works, congresses and research projects, after which we index all the terms, totaling and assigning a sequential integer value for each word. This equals more than 307,000 resumes and more than 1 million terms found. Then we eliminate less significant words, such as empty words, apply derivations and extract the radical from the word. By grouping the equal terms, we total 68,417 radicals. Each expert is now represented by a vector of size 68,417, and each value is calculated using TF-IDF, which takes into account the frequency of the term and how common it is, giving more value to the less used term. Finally, this data is used to train an Autoencoder and, as a result of an entry, it is reduced to 25 dimensions, before which it is expanded again, to minimize the reconstruction error. With this new data, we can use the same encoder to transform the queries in this 25-dimensional space and thus calculate the distance between the experts and the query.

Keywords: *Expert recommendation. Lattes platform. Neural Networks Deep Learning*

Una estrategia para la recomendación de expertos a partir de datos abiertos disponibles en la plataforma Lattes

RESUMEN

Con el creciente volumen de datos, es cada vez más difícil encontrar la información que desea. En este contexto, los agentes de recomendación y las técnicas de clasificación se consideran buenas soluciones para el problema de sobrecarga de información.Los xmls de los currículos son extraídos. Poco después, seleccionan-se los datos, extrayendo títulos de artículos de revistas, trabajos de anales, congresos y proyectos de investigación, después de lo cual indexamos todos los términos, totalizando y asignando un valor entero secuencial para cada palabra. Esto equivale a más de 307,000 hojas de vida y más de 1 millón de términos encontrados. Luego eliminamos palabras menos significativas, como palabras vacías, aplicamos derivaciones y extraemos el radical de la palabra. Al agrupar los términos iguales, totalizamos 68,417 radicales. Cada experto ahora está representado por un vector de tamaño 68,417, y cada valor se calcula utilizando TF-IDF, que tiene en cuenta la frecuencia del término y cuán común es, dando más valor al término menos utilizado. Finalmente, estos datos se usan para entrenar un Autoencoder y, como resultado de una entrada, se reduce a 25 dimensiones, antes de lo cual se amplía nuevamente, para minimizar el error de reconstrucción.Con estos nuevos datos, podemos usar el mismo codificador para transformar las consultas en este espacio de 25 dimensiones y así calcular la distancia entre los expertos y la consulta.

Palabras clave: *Recomendación de expertos. Plataforma Lattes. Redes neuronales. Deep Learning.*