Investigação de estratégias para utilização apenas da fase não supervisionada em desambiguadores híbridos de auto-treinamento

GUILHERME APARECIDO GREGORIO (Autor), Anderson Almeida Ferreira (Orientador), Guilherme Tavares de Assis (Co-Orientador)

Com o crescente número de dados sobre publicações científicas em bibliotecas digitais e sobre outros tipos de produção disponíveis em diversos meios,há a necessidade de que esses dados sejam de qualidade e essa qualidade pode ser prejudicada pela ambiguidade existente entre nomes de pessoas.Essa ambiguidade ocorre devido ao fato de um nome referenciar pessoas distintas ou nomes distintos referenciar a mesma pessoa.Esse problema prejudica tanto a recuperação de informação em um repositório de dados quanto o levantamento de informação sobre pessoas para tomada de decisão.Vale ressaltar que mesmo uma pessoa tendo um identificador único atualmente, como o ORCID no caso de autores, o legado ainda precisa ser resolvido e o mesmo estudo aqui pode se aplicar a outros contextos. Os algoritmos mais promissores para resolver esse problema são algoritmos que utilizam técnicas supervisionas de aprendizado, no entanto, normalmente necessitam de exemplos rotulados manualmente.Visando minimizar essa rotulagem manual, há propostas de algoritmos que tentam rotular automaticamente os exemplos por meio de aplicação de técnicas não supervisionadas e posteriormente a supervisionada. No entanto, verificou-se que, em algumas situações, apenas a fase não supervisionada seria suficiente. Assim, este trabalho teve como objetivo investigar e descobrir situações em que apenas a fase não supervisionada seria suficiente, evitando todo o custo da fase supervisionada desses algoritmos.Para isso,foram investigadas a aplicação de diversas métricas intrínsecas de avaliação de agrupamentos, coeficiente de Silhoetta (CS) , index DB, index CH, aos resultados das fases de um algoritmo de desambiguação,SAND, em dois conjuntos de dados de publicações científicas e um de patentes.Observou-se nos experimentos que para valores de CS acima de 0,848, em todos os conjuntos de dados, houve ganho na eficácia, sem a necessidade de executar a fase supervisionada nessa situação, comparada a execução completa do S AND.

Instituição de Ensino: Universidade Federal de Ouro Preto

ISSN: 21763410