O uso de anotação semântica e ontologias na busca de similaridade entre entrevistas não-estruturadas em banco de dados

Rovilson de Freitas[1], Elaine Parros Machado de Sousa[2]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (ICMC-USP) - Avenida Trabalhador São-carlense, 400 - Centro
CEP: 13566-590 - São Carlos - SP

[2]Departamento de Ciências da Computação - Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
(ICMC-USP) – São Carlos

rovilson.freitas@usp.br, parros@icmc.usp.br

Abstract. *The Olympic Studies Group (GEO), from the School of Education of the University of São Paulo (FE-USP), has carried out nearly twenty years of research related to Olympism. One of the research lines involves interviewing Brazilian Olympic athletes in an unstructured manner. There is a hypothesis that there are knowledge and similarities in this collection, but researchers need a computational tool to point them out. This work aims to propose a tool, based on semantic annotation and ontology, to meet this demand, not only assisting GEO's researchers, but also allowing other groups to benefit.*

Resumo. *O Grupo de Estudos Olímpicos (GEO), da Faculdade de Educação da Universidade de São Paulo (FE-USP), realiza há quase vinte anos diversas pesquisas relacionadas ao olimpismo. Uma delas envolve entrevistar, de maneira não-estruturada, os atletas olímpicos brasileiros. Existe a hipótese, por parte do grupo, de que há conhecimento e similaridades nesse acervo, mas os pesquisadores precisam de uma ferramenta computacional para aponta-los. O objetivo desse trabalho é justamente propor uma técnica que, baseada em anotação semântica e ontologia, possa atender essa demanda, não só apoiando o trabalho dos pesquisadores do GEO, mas também abrindo a possibilidade de que outros grupos sejam beneficiados.*

1. Introdução

Com aproximadamente vinte anos de existência, o Grupo de Estudos Olímpicos (GEO), da Faculdade de Educação da Universidade de São Paulo, atua em pesquisas relacionadas ao olimpismo. Uma de suas principais fontes de dados são as entrevistas, realizadas principalmente com atletas olímpicos. Essas entrevistas não possuem um roteiro definido, nem uma sequência pré-determinada. Também não estão armazenadas ou organizadas num sistema informatizado. Isso traz grandes dificuldades ao grupo, desde a busca por uma entrevista em específico até buscar possíveis objetos de pesquisa nelas. Muitas vezes, a análise de uma ou mais entrevistas dentro de um contexto pode proporcionar a descoberta de novos conhecimentos, resultando em novas pesquisas.

2. Objetivo

Partindo da necessidade de organização dessas entrevistas em meio informatizado, além de possivelmente facilitar o processo de análise por parte do pesquisador, o objetivo desse trabalho é criar uma técnica para encontrar a similaridade entre as entrevistas não-estruturadas, armazenadas num banco de dados relacional, usando os recursos de anotação semântica e ontologia. Para isso, o trabalho propõe a criação de uma ontologia para o Grupo de Estudos Olímpicos, considerando sua realidade e especificidade.

3. Ontologia em Ciências da Computação e Anotações Semânticas

No contexto das Ciências da Computação, Gruber (1995) define ontologia como uma especificação explícita de uma conceitualização. Essa modelagem utiliza o conceito de classes, atributos e relacionamentos. Essas informações reúnem dados sobre seus significados, restrições e aplicações lógicas existentes. Normalmente, são desenvolvidas numa linguagem que permite que a abstração da estrutura de dados e estratégias de implementação.

Guizzardi (2000) apud Guarino (1998), onde o autor estende a definição dada por Gruber:

> "()uma ontologia é na verdade uma especificação parcial e explícita que tenta, da melhor forma possível, aproximar a estrutura de mundo definida por uma conceituação. Uma ontologia, portanto, passa a ter compromisso apenas com a consistência em um determinado domínio e não com a completude. Ao conjunto de elementos de um domínio que podem ser representados em uma ontologia é dado o nome de universo de discurso."

Para Vickery (1997) e Smith (2004), ontologias podem alimentar um banco de dados com informações sobre categorias (conceitos) existentes no mundo/domínio, além das propriedades referentes a esses conceitos, bem como as relações existentes entre eles, além de permitir a integração de bancos de dados, de softwares ou de modelos de negócio.

Um outro conceito que, combinado com a ontologia pode auxiliar no processo de busca de similaridade, é o de anotação semântica.Esse processo, vincula modelos semânticos e linguagem natural, criando inter-relações entre ontologias e documentos estruturados ou não-estruturados (ARRUDA, 2017). Ainda segundo Arruda apud Li, Bontcheva (2007) e Kiriakov(2004):

> "Cria inter-relações entre ontologias e documentos não estruturados ou semiestruturados. A anotação semântica é a atribuição de links para a descrição semântica de cada entidade localizada nos documentos"

4. Trabalhos correlatos

Arruda (2017) apresenta em sua dissertação de mestrado a proposta de um método semântico baseado em ontologia (SOM4SImD) para detectar similaridade entre documentos no contexto da educação especial. Os resultados finais mostraram que o método SOM4SImD é mais vantajoso na obtenção de similaridade entre documentos. A título de comparação, esse trabalho teve um índice de precisão de 0,96, contra 0,71 de outro trabalho com características equivalentes.

Taieb, Aouicha e Hamadou (2014) apresentam uma nova medida para quantificar o grau de similaridade semântica entre conceitos e palavras com base na hierarquia WordNet e usando uma série de parâmetros topológicos relacionados à taxonomia "é um". Os resultados demonstram que, em comparação com outros métodos computacionais disponíveis atualmente, a medida apresentada nesse estudo resulta em melhores níveis de desempenho.

Mendonça e Soares (2017) propõem a aplicação da metodologia Ontoforinfoscience na elaboração de duas ontologias: a Hemonto, ontologia biomédica para componentes de sangue humano e a Ontolegis, uma ontologia de domínio jurídico. O trabalho conclui que, a metodologia foi útil para o desenvolvimento das ontologias, e seus detalhamentos permitiram auxiliar os desenvolvedores em questões lógicas e filosóficas do processo de construção e no entendimento de conceitos técnicos de ontologias.

Foi encontrado no dbpedia, um exemplo de ontologia no contexto de Jogos Olímpicos[1]. Essa ontologia, pode oferecer algumas informações importantes para a elaboração da ontologia do Grupo de Estudos Olímpicos.

5. Materiais e Métodos

O acervo de entrevistas do GEO será armazenado em uma base de dados criada utilizando o Sistema Gerenciador de Banco de Dados (SGBD) PostgreSQL[2]. Além de ser um sistema de código aberto (o que não trará custos para o grupo), o PostgreSQL oferece o suporte adequado para os tipos de dados envolvidos no projeto, incluindo textos longos (caso das entrevistas).

Neste trabalho, será desenvolvida uma ontologia, que terá como contexto principal alguns dos principais assuntos propostos e trabalhados pelo GEO. O Grupo traz algumas informações passiveis de futuros estudos, e esses dados formarão essa primeira versão da ontologia, que servirá de ponto inicial para as análises. Será utilizada, para o desenvolvimento dessa ontologia, a metodologia Ontoforinfoscience, proposta por Mendonça (2015). Para avaliar a sua eficiência, a própria metodologia propõe formas de avaliação do processo, divididas em duas partes: critério de validação (Compromisso ontológico, especificação, validação especializada e expansibilidade) e critério de verificação (Completude, integridade, consistência, precisão e documentação)

Uma vez armazenadas no banco de dados, será realizado o processo de anotação semântica nas entrevistas. Os termos utilizados na anotação serão fornecidos pela ontologia, considerando seus sinônimos e variações. Será utilizada a ferramenta GATE[3] como suporte para as anotações.

Para demonstrar o grau de similaridade entre as entrevistas, serão utilizadas as medidas de similaridade semântica de Lin (1998). Como ferramenta de suporte inicial, a proposta é

[1] http://dbpedia.org/ontology/olympicGames
[2] https://www.postgresql.org/

[3] https://gate.ac.uk/download/

utilizar a linguagem de programação Python[1], que oferece suporte para tarefas que envolvam alta capacidade de texto.

6. Considerações finais

Considerando as pesquisas conduzidas no Brasil nesse momento, em muitos casos o uso de entrevistas é fundamental. Transformar esse grande acervo de texto em resultados práticos, sem auxílio computacional, pode se tornar uma tarefa muito árdua. Oferecer aos pesquisadores uma ferramenta que possa auxiliá-los nesse processo, além de, potencialmente, acelerar o trabalho de pesquisa, também pode proporcionar que esses pesquisadores observem questões que antes não haviam sido percebidas.

Num contexto de entrevistas sem organização e perguntas pré-definidas, as dificuldades são ainda maiores. Afinal, não existe uma lógica que possa ser captada ou percebida de maneira mais simples, necessitando então, de uma técnica que possa mostrar o que é similar dentro delas e, eventualmente, até mesmo o que as diferencia.

Ao citar as entrevistas não-estruturadas (em especial no caso do GEO), é importante salientar que não se trata de uma desorganização. A abordagem do GEO para a realização das entrevistas está baseada em teorias notórias e importantes da área de humanidades (como narrativas biográficas, histórias de vida, preservação da memória, etc.). A entrevista nesse formato é fundamental e necessária para que os pesquisadores do grupo possam conduzir seus trabalhos de acordo com essas teorias.

Ainda assim, mediante a todos os recursos escassos da realidade atual, é necessário que esses pesquisadores possam ter um suporte computacional, para que esse trabalho seja realizado mais rapidamente, evitando ações repetitivas e/ou desnecessárias.

Com esse projeto, pretende-se mostrar que o uso de ontologias e anotações semânticas pode gerar uma técnica e, consequentemente, uma ferramenta que ajude nesse processo, beneficiando não apenas esse grupo em especial, mas outros grupos de estudo que tem o mesmo procedimento em suas pesquisas, contribuindo para futuras descobertas no campo científico.

Referências

ARRUDA, C. G. d. SOM4SImD: Um método semântico baseado em ontologia para detectar similaridade entre documentos. Dissertação (Mestrado em Ciência da Computação) - Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos, São Carlos, 2017. Disponível em: <https://repositorio.ufscar.br/handle/ufscar/8961> Acesso em: 12-05-2019.

CASTRO, S. Ontologia. Rio de Janeiro: Jorge Zahar, 2008

GRUBER, T. R. Toward Principles for the Design of Ontologies used for Knowledge Sharing. Int. J. Human-Computer Studies, v. 43, n. 5/6, 1995. Disponível em: < https://www.sciencedirect.com/science/article/abs/pii/S1071581985710816> Acesso em: 18-08-2020.

GUIZZARDI, G. Uma abordagem metodológica de desenvolvimento para e com reuso, baseada em ontologias formais de domínio, Dissertação (Mestrado em Ciências da

[1] https://www.python.org/

Computação) Universidade Federal do Espírito Santo, Brasil. 2000. Disponível em < http://www.inf.ufes.br/~gguizzardi/dissertacao_msc.pdf> Acessado em: 20-08-2020.

LIN, D. An information-theoretic definition of similarity. In Proceedings of the 15th International Conference on Machine Learning, Madison, WI, 1998. Disponível em: < https://www.cse.iitb.ac.in/~cs626-449/Papers/WordSimilarity/3.pdf>. Acesso em 15-05-2019.

MENDONÇA, F. M. OntoForInfoScience: metodologia para construção de ontologias pelos cientistas da informação: uma aplicação prática no desenvolvimento da ontologia sobre componentes do sangue humano (Hemonto). 2015. Tese (Doutorado)-Universidade Federal de Minas Gerais, Belo Horizonte, Brasil, 2015.

MENDONCA F. M.; SOARES A. L. Construindo ontologias com a metodologia ontoforinfoscience: uma abordagem detalhada das atividades do desenvolvimento ontológico. Ciência da Informação, v. 46, n. 1, 28 dez. 2017. Disponível em: <http://revista.ibict.br/ciinf/article/view/4013/3713> Acesso em 07-10/2020.

SMITH, B. Ontology and Information Systems, 2004. Disponível em: <http://www.ontology.buffalo.edu/ontology> Acesso em: 20-07-2019.

TAIEB, M.A.H; AOUICHA, M.B; HAMADOU, A.B. Ontology-based approach for measuring semantic similarity, Engineering Applications of Artificial Intelligence, Volume 36, 2014, pg 238-261, Disponível em <http://www.sciencedirect.com/science/article/pii/S0952197614001833>.

VICKERY, B. C. Ontologies. Journal of Information Science, v. 23, n. 4, p. 277-286, 1997. Disponível em: <https://journals.sagepub.com/doi/10.1177/016555159702300402>. Acesso em: 17-07-2019.