UM PROCESSO PARA APOIAR A DESAMBIGUAÇÃO DE NOMES EM CURRÍCULOS LATTES BASEADO EM WEB SEMÂNTICA

ANDRÉ GUIMARÃES PEIL[1]; DANIELA FRANCISCO BRAUNER[2], RICARDO MATSUMURA DE ARAÚJO[3];

[1]*Universidade Federal de Pelotas – agpeil@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – danibrauner@inf.ufpel.edu.br*
[3]*Universidade Federal de Pelotas – ricardo@inf.ufpel.edu.br*

1. INTRODUÇÃO

A Plataforma Lattes, implementada e mantida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), é o principal meio de exposição de trabalhos científicos dos pesquisadores brasileiros. Por conter um grande volume de dados, que abrange perfis de pesquisadores, estudantes, linhas de pesquisa, projetos, publicações entre outros assuntos, é possível reconhecer alguns problemas ao analisar certos perfis, como a possibilidade de ambiguidades nos registros de publicações. Como a maioria dos pesquisadores não registra seus co-autores com os devidos links para seus currículos, a citação atribuída aos co-autores torna-se uma simples cadeia de caracteres (*string*) e pode causar problemas de ambiguidade quando por exemplo, temos pessoas que possuem mesmo sobrenome e iniciais iguais, impossibilitando um terceiro de reconhecer a qual pesquisador se refere tal co-autoria. Neste contexto, este trabalho abordará o desenvolvimento de um processo para identificar co-autores dentre os pesquisadores de um grupo de currículos Lattes.

A Web Semântica (Berners-Lee, T., 1990) é uma extensão da Web atual onde a informação recebe um significado bem definido, permitindo melhor interação entre os computadores e as pessoas. As ontologias[1] são utilizadas para anotar semanticamente o conteúdo disponibilizado, o que permite que agentes de software compreendam a semântica embutida nas páginas Web, sem ambiguidade, viabilizando o intercâmbio de informações. A RDF (*Resource Description Framework*) é uma linguagem de propósito geral baseada em XML, utilizada para representar ontologias e informações na Web. É possível expressar em RDF o significado de conceitos, propriedades e valores. Um dos formatos de apresentação dos dados em RDF é o formato de triplas, compostas por: sujeito, predicado e objeto. Neste trabalho, é proposto o uso de RDF para representar os currículos dos pesquisadores. Os dados são importados do currículo Lattes, transformados para RDF e armazenados em formato de triplas numa base de dados específica para armazenamento de ontologias, que será apresentada mais adiante.

Este trabalho tem como objetivo combinar técnicas e ferramentas baseadas em Web Semântica para propor um processo de apoio a desambiguação de nomes de co-autores em currículos de pesquisadores. Além disso, visa contribuir para o projeto Europa[2], desenvolvido pela Empresa Jr. da Computação (Hut8) no sentido de fornecer uma alternativa para tratar as informações advindas da Plataforma Lattes (PL).

---

[1] Ontologia: uma especificação de uma conceitualização, ou seja, é uma descrição de conceitos e relacionamentos que existem entre estes conceitos (Gruber, 1992).

[2] Projeto Europa é um sistema web que permite fazer análise gráfica e numérica de currículos Lattes.

2. METODOLOGIA

A metodologia utilizada para alcance dos resultados deste trabalho baseou-se nas seguintes etapas: entendimento da Plataforma Lattes e seu formato de exportação, revisão bibliográfica sobre desambiguação de nomes e as técnicas propostas, desenho do processo, implementação e testes.

Esse método foi realizado de forma interativa e incremental, permitindo que, a cada novo ciclo, sejam desenvolvidas novas tarefas. Por exemplo, os testes foram alimentando melhorias no desenho do processo.

3. RESULTADOS E DISCUSSÃO

Inicialmente foi feito um estudo para entendimento dos detalhes e as peculiaridades da Plataforma Lattes e de seu formato de exportação de dados: XML (eXtensible Markup Language). Cada instituição de ensino superior e pesquisa brasileira tem acesso a um sistema, disponibilizado pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), o mesmo mantenedor da Plataforma Lattes, para download dos dados do Lattes dos pesquisadores de sua instituição. O sistema disponibiliza os dados em XML. Com acesso a este sistema, foram baixados mais de 1000 currículos dos pesquisadores da UFPel. Porém, como forma de dar semântica a estas informações, como parte do processo a ser proposto neste trabalho, de posse destes dados em XML, utilizamos o SLATTES[3], um script para transformação dos dados de currículos Lattes para dados anotados seguindo a ontologia VIVO-ISF em RDF. A VIVO-ISF[4] é uma ontologia para representar informações sobre pesquisadores, publicações, projetos e outras. Os dados foram então armazenados na base OpenLink Virtuoso[5], que é um servidor universal que oferece, dentre outras funcionalidades, armazenamento de dados, servidor de aplicação Web e *triple store*, i.e., uma ferramenta para armazenamento e consultas de dados em RDF (triplas RDF). Ele fornece um mecanismo para armazenamento persistente e acesso à grafos RDF. O Virtuoso oferece uma interface de consultas SPARQL, que é a linguagem de consultas utilizada para obter dados de RDF. Para este trabalho, foi utilizada a versão gratuita do OpenLink Virtuoso, para servir de armazenamento dos currículos em RDF que também permite acessar as informações em formato de triplas[6].

Em seguida, a revisão bibliográfica abrangeu conceitos de Linked data, Web semântica e desambiguação de nomes. Métodos tradicionais de desambiguação comparam a informação sintática dos atributos dos objetos ambíguos e, utilizando funções de similaridade e outras heurísticas, determinam se estes objetos representam a mesma entidade real ou não (LEVIN,2010).

Neste caminho, durante a revisão bibliográfica encontramos algumas abordagens importantes e que fundamentam tanto na teoria quanto na prática a forma que é organizada a estrutura de desambiguação de nomes. A abordagem baseada em aprendizado de máquina proposta por TORVIK (2009), agrupa um conjunto de obras de literatura, correspondentes às pessoas que as escreveram.

---

[3] https://github.com/arademaker/SLattes/

[4] http://issues.library.cornell.edu/browse/VIVOONT

[5] http://virtuoso.openlinksw.com

[6] Para DA SILVA; LIMA (2001), uma tripla é formada por um recurso, uma propriedade e um valor para a propriedade daquele recurso. Possui o seguinte formato <sujeito, predicado, objeto>

Geralmente esta técnica requer a aquisição de conjuntos de treinamento que fornecem exemplos de falsos positivos e negativos; a extração de uma ou mais características das obras ou dos metadados, além de recursos de otimização ou aprendizagem, que atuam de acordo com as características citadas.

Outra abordagem interessante é a de BHATTACHARYA; GETOOR (2007), onde é proposta uma técnica de desambiguação de entidades comparando-as coletivamente. Trata-se de um algoritmo que utiliza atributos como nome, título e também relacionamentos entre co-autores para aplicar o processo de desambiguação. Além disso, um fator importante nesta técnica é a geração de um valor de similaridade de vizinhança, onde os autores, para serem considerados duplicatas, precisam ter um conjunto semelhante de co-autores para firmar a correspondência. Portanto, ao invés de comparar sempre pares de entidades dois a dois, o algoritmo compara grupos de objetos duplicados, (porém na fase inicial, estes grupos são unitários) e, à medida em que os grupos casam, os objetos de ambos os grupos representam a mesma entidade, ou seja, os grupos vão sendo unidos. Ambas as abordagens citadas tiveram influência no desenho do processo proposto, como poderá ser visto na próxima sessão.

De posse do conjunto de dados em RDF armazenado na base de dados Virtuoso e após a revisão bibliográfica, o processo de desambiguação foi enfim proposto, implementado e testado de forma incremental. O processo proposto neste trabalho foi desenhado para conter três etapas que seguem: Normalização, Clusterização e Pareamento. A Figura 1 apresenta uma representação do processo de desambiguação.

Figura 1: Processo para desambiguação de nomes

Na etapa de normalização são retirados todos os tipos de caracteres que podem influenciar na diferenciação das *strings* (as strings normalizadas são os nomes dos autores, título de artigo e congresso), para obter melhores resultados são utilizados alguns métodos que auxiliam neste processo como retirar acentos, *stopwords*[7], conectivos, substituir letras maiúsculas por minúsculas, caracteres e letras especiais como por exemplo o 'ç'.

Na segunda etapa é feita a clusterização baseada na abordagem de *Semantic Graph Blocking* (SGB) é uma técnica apresentada por (NIN, et al., 2007) que propõe uma nova família de algoritmos de blocagem que constroem blocos baseados no contexto. A finalidade da blocagem em deduplicação de informações é através de algum critério, agrupar os dados em conjuntos de dados semelhantes entre si, com o intuito de restringir o número de comparações entre objetos, LEVIN, (2010). No escopo deste trabalho o critério para a blocagem é o nome do artigo.

Na terceira e última etapa é feito o pareamento dos itens dos blocos, ou seja, cada bloco possui todas as instâncias de co-autores que podem ser considerados semelhantes, neste caso é comparado um a um, até que se chegue em um peso de similaridade entre pares de instâncias, a Figura 2 demostra a

[7] stopwords: palavras irrelevantes. ÁLVAREZ, (2007).

representação de co-autores que estão em diferentes perfis mas que representam a mesma pessoa.

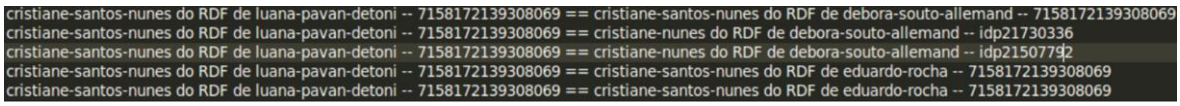

Figura 2: Log de desambiguação Fonte: https://github.com/andreguipeil/evo

4. CONCLUSÕES

Neste trabalho foi apresentado um processo que combina o uso de técnicas e ferramentas baseadas em Web Semântica para contribuir com a questão de desambiguação de nomes de co-autores em currículos de pesquisadores. Além disso, visa contribuir para o projeto Europa, no sentido de fornecer uma alternativa para tratar as informações advindas da Plataforma Lattes (PL).

Este trabalho é parte de um trabalho de conclusão de curso que dentre seus objetivos inclui a criação de uma forma de validar o resultado e representar a informação, baseada em ontologias. A ideia é criar uma interface que permita ao usuário manipular um limiar de desambiguação, por exemplo, criar um nível fraco e forte que representaria o nível de ambiguidade que deseja visualizar a informação. Além disso, as próximas atividades incluem a criação de regras baseadas em ontologias que dêem suporte ao resultado da desambiguação e a interface que permita o usuário interagir com a aplicação.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ÁLVAREZ, A. C. **Extração de infomação de artigos científicos: uma abordagem baseada em indução de regras de etiquetagem**. 2007. Dissertação de Mestrado Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos/SP.

BERNERS-LEE, T., HENDLER, J. and LASSILA, O. (2001). **The Semantic Web**. Scientific American, 2 84(5), 35–43.

BHATTACHARYA, I.; GETOOR, L. 2007. **Collective entity resolution in relational data**. *ACM Trans. Knowl. Discov. Data* 1, 1, Article 5 (March 2007). DOI=10.1145/1217299.1217304.

GRUBER, T. **What is an ontology** (1992). Disponível em: <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html>, Acessado em 9 de julho de 2015.

LEVIN, F H. **Desambiguação de Autores em Bibliotecas Digitais utilizando Redes Sociais e Programação Genética**. 2010. Dissertação de Mestrado PPGC/UFGRS, Porto Alegre/RS

NIN, J.; MUNTÉS-MULERO, V.; MARTINEZ-BAZAN, N.; LARRIBA-PEY, J. On the Use of Semantic Blocking Techniques for Data Cleansing and Integration. In: **INTERNATIONAL DATABASE ENGINEERING AND APPLICATIONS SYMPOSIUM**, 11., Banff, Canadá, 2007. Proceedings... Washington, DC, EUA: IEEE Computer Society, 2007, p.190-198.

SILVA, G. C. da; SOUZA LIMA, T. de. **RDF e RDFS na Infra-estrutura de Suporte à Web Semântica**. Acessado em 15 jun. 2015. Disponível em: www2.ic.uff.br/~gsilva/slreic.pdf

TORVIK, V. **Author Name Disambiguation in MEDLINE. ACM transactions on knowledge discovery from data**, [S.l.], p.5 – 19, 6 2009.