Integração semântica entre dados dos domínios da educação e segurança: um caso de Curitiba

Pedro Henrique Stolarski Auceli[1], Rita C. G. Berardi[1] Nadia P. Kozievitch[1]

Departamento Acadêmico de Informática
Universidade Tecnológica Federal do Paraná (UTFPR) – Curitiba, PR – Brazil
pedroauceli@gmail.com, ritaberardi@utfpr.edu.br, nadiap@utfpr.edu.br

Abstract. *The objective of this work is to analyze if there is a relationship between the students' income and the number of police occurrences in the neighborhood in which they are based on the semantic integration of heterogeneous open databases of education and security domains. To perform the integration an ontology is proposed with the intention of semantically unify the base domain and formally specify the data relationship.*

Resumo. *O objetivo desse trabalho é analisar se existe uma relação entre o rendimento de alunos com a quantidade de ocorrências policiais do bairro em que se encontram a partir da integração semântica de bases de dados abertas heterogêneas dos domínios de educação e segurança. Os dados utilizados são referentes à cidade de Curitiba-Paraná. Para realizar a integração uma ontologia é proposta com o intuito de unificar semanticamente o domínio das bases e especificar formalmente o relacionamento dos dados.*

1. Introdução

Atualmente no Brasil existem várias bases de dados abertas e, pelo fato de cada uma ser de um domínio específico diferente, realizar uma integração entre elas para recuperar informações úteis e mais complexas é uma tarefa não trivial. Isso se deve às diferentes semânticas dos dados, ou seja, a diferença do significado dos dados nas diferentes bases, que são modeladas, coletadas e abertas de maneira independente.

O objetivo desse trabalho é a integração semântica de bases de dados, assim possibilitando novos tipos de análises sobre os dados. Para alcançar tal objetivo é criada uma ontologia, que é uma especificação de uma realidade (GUARINO; OBERLE; STAAB, 2009), que pode ser utilizada para integrar bases de dados de domínios diferentes, uma vez que nela será especificado um novo domínio que unifica as bases. No caso deste trabalho as bases de dados são da área da educação e da segurança pública. As bases foram escolhidas a partir do argumento de Junior et al. (2018), que diz que trabalhos com dados abertos governamentais devem ser feitos com dados que sejam relevantes para a população, com o intuito de melhorar os serviços ofertados pelo governo. O método apresentado por Pereira, Salvador, Wassermann (2018) será utilizado para avaliar a ontologia. Esse consiste em criar uma pergunta que deve ser respondida com informações apenas obtidas através dos dados integrados (PEREIRA, SALVADOR, WASSERMANN, 2018). No caso deste trabalho a pergunta feita é: existe uma relação entre o rendimento e as notas de escolas com a quantidade de ocorrências policiais do bairro em que se encontram?

2. Bases de dados

As bases de dados que são utilizadas neste trabalho são referentes à cidade de Curitiba: SiGesGuarda e Unidades de Atendimento de Curitiba ativas, além dos datasets de Média e Rendimento dos alunos por região. Tanto a base de nota média de escolas quanto a de rendimentos foram obtidas através do portal de dados abertos do governo brasileiro[1], e contêm respectivamente as notas de acordo com as turmas de cada escola dentro do país e o rendimento das mesmas. O rendimento é um cálculo feito através da taxa de aprovação, reprovação e abandono dos alunos. Ambas as bases se encontram no formato xls (formato proprietário do Microsoft Excel) e podem ser convertidas para csv (*Comma-separated Values*). A base da SiGesGuarda é referente aos dados de atendimentos feitos pela guarda municipal da cidade de Curitiba, que pode ser obtida em formato csv através do portal de dados abertos da cidade de Curitiba[2]. A base de unidades de atendimento ativas também é disponibilizada através do portal de dados abertos de Curitiba, e é referente às unidades de atendimento de uso público.

3. Metodologia

O primeiro passo para conseguir integrar as bases foi fazer a limpeza, a normalização e redução da granularidade dos dados de cada uma delas. O tratamento dos dados é necessário para facilitar a comparação, e para poder inserir os dados dentro de um banco de dados relacional. Para utilizar o plugin Ontop, utilizado no framework Ontop apresentado por Pereira, Salvador, Wassermann (2018), o banco de dados relacional é necessário, pois ele é o responsável pela distribuição dos dados que servirão para povoar a ontologia. Enquanto que a redução foi feita para minimizar o custo computacional e diminuir a complexidade da especificação da ontologia. O passo seguinte foi a inserção dos dados obtidos através das bases de dados no banco de dados PostgreSQL. Vale ressaltar que foram utilizados apenas 500 registros da base da SiGesGuarda, e que os dados das outras 3 bases foram inseridos manualmente pelo autor, em uma quantidade suficiente para testar a ontologia. Isso ocorreu devido à grande quantidade de ruídos que dificultaram o trabalho de limpeza e inserção dos dados.

Utilizando a ferramenta Protégé[3], foi criada a ontologia (Figura 2) com suas classes, relacionamentos e propriedades necessárias para responder à questão de competência motivadora ao experimento. No total foram criadas 3 classes: Bairro, GuardaMunicipal e Escola. 1 relacionamento: hasBairro, que liga um registro da classe GuardaMunicipal ou Escola com um bairro. E 8 propriedades de dados: nomeEscola, nomeBairro que são do tipo *String,* mesRegistro, anoRegistro, codigoRegistro, codigoBairro que são do tipo *int* e mediaEscola, rendimentoEscola que são do tipo *float*.

Com a ontologia devidamente criada foi necessário definir as regras para o mapeamento do banco de dados relacional. Na Figura 1 são apresentados todos os mapeamentos que foram necessários entre a ontologia e o banco de dados relacional.

---

[1] Governo Brasileiro. Portal brasileiro de dados abertos.< http://dados.gov.br/group/educacao>

[2] Prefeitura de Curitiba. Portal de dados abertos da cidade de Curitiba. <http://www.curitiba.pr.gov.br/dadosabertos/consulta/>

[3] The Protégé project: <https://protege.stanford.edu/>

4. Resultados

Com a ontologia povoada a integração foi obtida e sua avaliação pode ser realizada. Para avaliar se a ontologia foi suficiente para a integração, será utilizada a questão de competência apresentada na introdução. Para isso foi utilizado outro plugin chamado Ontop SPARQL, que permite a criação de *queries* em SPARQL que serão executadas sobre a ontologia.

Figura 1. Mapeamentos entre a ontologia e o banco de dados relacional

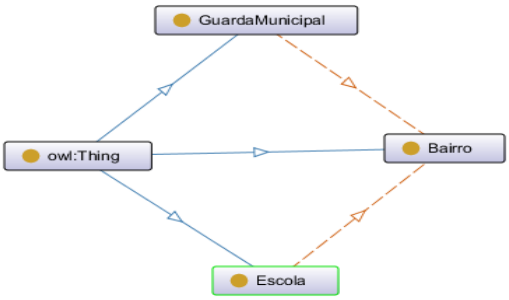

Figura 2. Ontologia

Uma dificuldade encontrada foi o fato da função *GROUP BY* não ter sido implementada pela equipe do Ontop, logo não foi possível utilizar a função *count* e não foi possível encontrar o Bairro com a maior quantidade de registros de atendimentos. Para resolver isso foram feitas *queries* específicas para cada bairro como pode ser visto na query abaixo, em que é tratado especificamente o bairro Bacacheri da cidade de Curitiba.

```
PREFIX tes: http://example.org/
SELECT ?registro ?nome ?escola ?media ?rendimento
WHERE {
?bairro tes:nomeBairro ?nome.
?bairro tes:nomeBairro "bacacheri".
?registro tes:hasBairro ?bairro.
?registro a tes:GuardaMunicipal.
?escola tes:hasBairro ?bairro.
?escola a tes:Escola.
?escola tes:mediaEscola ?media.
?escola tes:rendimentoEscola ?rendimento
}
```

rendimento	escola	nome	media	registro
"98.0"^^xsd:double	<http://example.org/Rosario>	"bacacheri"	"92.0"^^xsd:double	<http://example.org/GuardaMuni...
"98.0"^^xsd:double	<http://example.org/Rosario>	"bacacheri"	"92.0"^^xsd:double	<http://example.org/GuardaMuni...
"98.0"^^xsd:double	<http://example.org/Rosario>	"bacacheri"	"92.0"^^xsd:double	<http://example.org/GuardaMuni...
"98.0"^^xsd:double	<http://example.org/Rosario>	"bacacheri"	"92.0"^^xsd:double	<http://example.org/GuardaMuni...
"98.0"^^xsd:double	<http://example.org/Rosario>	"bacacheri"	"92.0"^^xsd:double	<http://example.org/GuardaMuni...
"98.0"^^xsd:double	<http://example.org/Rosario>	"bacacheri"	"92.0"^^xsd:double	<http://example.org/GuardaMuni...
"98.0"^^xsd:double	<http://example.org/Rosario>	"bacacheri"	"92.0"^^xsd:double	<http://example.org/GuardaMuni...
"98.0"^^xsd:double	<http://example.org/Rosario>	"bacacheri"	"92.0"^^xsd:double	<http://example.org/GuardaMuni...
"98.0"^^xsd:double	<http://example.org/Rosario>	"bacacheri"	"92.0"^^xsd:double	<http://example.org/GuardaMuni...
"98.0"^^xsd:double	<http://example.org/Rosario>	"bacacheri"	"92.0"^^xsd:double	<http://example.org/GuardaMuni...
"98.0"^^xsd:double	<http://example.org/Rosario>	"bacacheri"	"92.0"^^xsd:double	<http://example.org/GuardaMuni...
"98.0"^^xsd:double	<http://example.org/Rosario>	"bacacheri"	"92.0"^^xsd:double	<http://example.org/GuardaMuni...
"98.0"^^xsd:double	<http://example.org/Rosario>	"bacacheri"	"92.0"^^xsd:double	<http://example.org/GuardaMuni...
"98.0"^^xsd:double	<http://example.org/Rosario>	"bacacheri"	"92.0"^^xsd:double	<http://example.org/GuardaMuni

Figura 3. Resultados da *query*

Na Figura 3 é apresentado o resultado da *query* com o bairro com a maior quantidade de atendimentos feitos pela guarda municipal. Onde nesse caso o bairro com a maior ocorrência foi o bairro "Bacacheri", e graças à integração é possível verificar a média e o rendimento das escolas da área. A Figura 3 deixa evidente as limitações da ferramenta, por não conseguir agrupar os registros, porém também mostra que é possível obter a integração e fazer uma análise dos dados.

5. Considerações finais

O experimento realizado neste trabalho teve resultados satisfatórios, uma vez que foi possível obter uma análise com base nos dados integrados. Porém vale ressaltar que a análise foi feita com uma porção dos dados disponíveis pelas bases de dados.

Foram observadas dificuldades com relação às bases de dados, uma vez que é necessária uma limpeza dos dados antes da utilização desses. Além disso, foram encontrados problemas na implementação, uma vez que algumas funções do plugin Ontop SPARQL, como a *GROUP BY* e a *count,* ainda não foram implementadas, o que dificultou a obtenção de resultados através da análise das *queries*. Outro problema encontrado foi a redundância no mapeamento de classes, principalmente para a classe "Bairro", onde novas instâncias eram criadas para cada registro das bases de dados que continham um bairro.A utilização da ontologia e dos mapeamentos entre a ontologia e os bancos de dados relacionais se mostrou promissora para a realização de uma integração entre os domínios heterogêneos. Nas próximas análises serão utilizados mais dados e novas perguntas serão executadas sobre os domínios integrados.

References

Guarino N., Oberle D., Staab S. (2009) "What Is an Ontology?". In: Staab S., Studer R. (eds) Handbook on Ontologies. International Handbooks on Information Systems. Springer, Berlin, Heidelberg

JUNIOR, F. T. M. et al. "Avaliação da prontidão para abertura de dados das instituições públicas brasileiras: caso de uma instituição financeira pública brasileira". Brazilian Journal of Information Studies: Research Trends, 2018.

Pereira, D. L. N. C., Wassermann, R., Salvador, L. Integração Semântica das Bases de Dados do Município de São Paulo: Um Estudo de Caso com Anomalias Congênitas. XI Seminar on Ontology Research in Brazil, ONTOBRAS 2018.

Framework Ontop. Disponível em: <https://ontop.inf.unibz.it/>

Musen, M.A. The Protégé project: A look back and a look forward. 2015. Association of Computing Machinery Specific Interest Group in Artificial Intelligence.