UNIVERSIDADE FEDERAL DE PERNAMBUCO

GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CENTRO DE INFORMÁTICA

2015.2

ESTUDO SOBRE FUNÇÕES DE SIMILARIDADES CONSIDERANDO ASPECTOS SEMÂNTICOS

PROPOSTA DE TRABALHO DE GRADUAÇÃO

Aluno: Fernando Maranhão Pessoa Nazareth (fmpn@cin.ufpe.br)
Orientador: Ana Carolina Salgado (acs@cin.ufpe.br)

Recife, Outubro de 2015

1. Resumo

Funções de similaridade são utilizadas para medir quanto um objeto ou um dado qualquer é semelhante ao outro, a partir de métricas que definem o grau de similaridade entre eles. Sua importância deve-se ao fato que essas funções representam uma importante etapa na comparação de pares de registros durante o processo de Detecção de Dados Duplicados. A maioria das funções de similaridades existentes trata apenas o aspecto sintático na comparação dos dados. O objetivo deste trabalho é identificar, classificar e mensurar essas funções considerando o aspecto semântico na comparação de registros de dados.

2. Contextualização

Dado o aumento considerável da coleção de dados gerados na última década, a necessidade de integrar esses dados de forma a aumentar a qualidade ou facilitar a compreensão de informações, expandiu o interesse por novas técnicas que possam eficientemente gerenciar e analisar grandes coleções de registros. Em geral, a integração de banco de dados se divide em três tarefas. A primeira tarefa é a Correspondência entre Esquemas, nela se identificam tabelas, atributos e estruturas conceituais correspondentes. A segunda etapa, Detecção de Dados Duplicados, consiste em identificar e associar registros individuais de diferentes bancos de dados que se referem ao mesmo objeto. A terceira e última tarefa refere-se à fusão de dados, este processo consiste em fundir pares de registros que tenham sido identificados como pertencentes a uma mesma entidade.

O presente trabalho está incluído na etapa Detecção de Dados Duplicados. Detectar dados duplicados consiste em encontrar registros que se referem a uma mesma entidade em uma ou mais fontes de dados. Fontes de dados podem conter registros duplicados de entidades do mundo real por diversas razões. Erros na entrada dos dados, não padronização de abreviações, pequenas diferenças nos esquemas em múltiplas

fontes de dados são apenas alguns dos motivos que ocasionam esse cenário.

A etapa de Detecção de Dados Duplicados pode ser segmentada em diversos passos, cujo objetivo final é preparar para o processo de fusão de dados. As etapas são por ordem: Pré-processamento de Dados, Indexação, Comparação entre Pares de Registros, Classificação de Pares de Registros, Avaliação da Qualidade e Complexidade da Correspondência.

O foco do presente projeto é identificar, classificar e mensurar funções de similaridade semântica que serão utilizadas na terceira etapa do processo de detecção de dados duplicados, a Comparação entre Pares de Registros. Funções de Similaridade são utilizadas para medir quanto um objeto ou um dado qualquer é similar ao outro, a partir de métricas que definem o grau de similaridade entre eles.

3. Objetivos

O objetivo geral desse Trabalho de Graduação é o estudo aprofundado das Funções de Similaridade considerando aspectos semânticos. Ele é parte integrante de um estudo mais amplo sobre Detecção de Dados Duplicados tema de um projeto de doutorado em andamento.

No processo da construção do Trabalho de Graduação, os algoritmos das funções de similaridade existentes serão identificados e analisados. Alguns deles serão implementados bem como serão realizados testes e avaliação de desempenho dos mesmos. Por fim, será elaborado o relatório final de forma a apresentar uma síntese de todo o esforço empregado durante o processo de realização deste trabalho.

4. Cronograma

Nesta seção, é apresentado o cronograma de atividades previsto para o desenvolvimento desse Trabalho de Graduação (Tabela 1).

Atividade	Setembro				Outubro				Novembro				Dezembro			
Levantamento do estado da arte e definição do escopo		X	X	X												
Escolha e Implementação dos algoritmos			X	X	X	X	X									
Testes e experimentos				X	X	X	X	X	X							
Análise dos resultados							X	X	X	X						
Elaboração do relatório				X	X	X	X	X	X	X	X	X	X			
Preparação e defesa												X	X	X		

Tabela 1: Cronograma de atividades

5. Possíveis Avaliadores

Os possíveis avaliadores para o resultado a ser obtido ao final de todas as etapas da proposta descrita neste documento são:

- Bernadette Farias Lóscio
- Fernando da Fonseca de Souza

6.Assinaturas

_____

Fernando Maranhão Pessoa Nazareth
Orientando

_____

Ana Carolina Salgado
Orientador

Recife, Outubro de 2015