Universidade Federal de Pernambuco
Graduação em Ciências da Computação
Centro de Informática
2015.1

Agrupamento de Instâncias no Processo de Identificação de Dados Duplicados

Recife, Abril de 2015
Proposta de trabalho de graduação

Aluno: Augusto Juvenal Fonseca Giles Costa (ajfgc@cin.ufpe.br)
Orientador: Ana Carolina Salgado (acs@cin.ufpe.br)

1. Resumo

Na década de 2000 houve uma crescente exponencial de dados, a qual preocupa analistas por uma futura falta de espaço de armazenamento. Segundo a IBM, 90% dos dados virtuais foram produzidos nos últimos dois anos da década citada, decorrente da inserção de grandes empresas à internet, como também da criação de várias redes sociais, dados de dispositivos móveis, GPS, entre outros. As grandes organizações agora buscam utilizar este grande volume de dados, voláteis ou não, em seu benefício. O desafio vem na forma de identificar sentimentos semelhantes (reclamações, elogios, críticas,...) dentro deste grande volume, como também identificar dados que pertencem a uma mesma entidade (pessoa, empresa, semáforo, ônibus,...).

O objetivo deste trabalho vem do desafio da recuperação de dados em diversas fontes de dados, com a motivação de integrar instâncias semelhantes por meio de agrupamento inteligente, este processo deve reconhecer dados passíveis de conterem algumas diferenças, como pertencentes a uma mesma entidade, agrupando-os em *clusters*.

2. Contextualização

A era Big Data levanta dois desafios para agrupamento de dados semelhantes: o volume de informação é gigante e a velocidade de atualizações de dados é muitas vezes elevada, tornando resultados anteriores obsoletos.

No cenário de pesquisas para a Internet das coisas, é viável prever que cada sistema de banco de dados crie seus próprios identificadores, sua estrutura e semântica que mais convém, dificultando ainda mais o processo de integração de diferentes fontes de dados. Nesse contexto, o gerenciamento de dados diante de um ambiente heterogêneo torna-se desafiador a fim de encontrar informações de interesse do usuário, localizando-os através de diversas fontes de dados, decidindo, dentre as chaves de identificação definidas, se as instâncias pertencem ou não a uma mesma entidade.

O agrupamento de dados tem função importante no processo de integração para colocar juntos dados que se assemelhem e representem a mesma entidade. Através de métodos, dependentes do algoritmo de agrupamento, é possível calcular uma margem

de semelhança entre cada entidade, sendo capaz de agrupar ou identificar diferentes menções de uma mesma entidade em fontes de dados diferentes.

3. Objetivos

O objetivo geral deste Trabalho de Graduação é a especificação e implementação de um algoritmo incremental de identificação de instâncias duplicadas (Agrupamento), analisando múltiplas fontes de dados em diferentes arquiteturas de gerenciamento de dados.

A abordagem incremental é necessária quando se trata de uma análise de dados da ordem de grandeza da Web, onde a quantidade de atualizações (inserções, remoções,...) acontece a cada segundo em diversas instâncias de informações, a análise anterior deve ser reaproveitada de algum modo para integrar os dados atualizados a fim de responder em tempo hábil a consultas em grande volumes de dados.

4. Cronograma

Atividade	Abril				Maio				Junho				Julho			
Levantamento inicial e definição do escopo		x	x	x												
Implementação do algoritmo					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	x	
Preparação da 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

Augusto Juvenal F. G. Costa
Orientando

Ana Carolina Salgado
Orientador

Recife, Abril de 2015