

## **AGRUPAMENTO DE REGISTROS DUPLICADOS NO APLICATIVO DEDUP**

**LINDENAU, Guilherme; MACHADO, Rafael; PINHEIRO, Rafael; NUNES, Eliza  
BORGES, Eduardo N.  
guilhermelindenau@furg.br**

**Evento: Congresso de Iniciação Científica**

**Área do conhecimento: Ciências Exatas e da Terra, Ciência da Computação**

**Palavras-chave:** agrupamento; identificação de duplicatas; Android

### **1 INTRODUÇÃO**

O aplicativo DeDup, desenvolvido pelos estudantes de IC do Grupo de Pesquisa em Gerenciamento de Informações da FURG, coleta os contatos do dispositivo Android no qual está instalado e identifica os registros duplicados, ou seja, contatos com diferentes nomes, e-mails ou telefones que representam a mesma pessoa. A concepção e arquitetura do aplicativo foram publicadas anteriormente (PINHEIRO et. al, 2014). Este trabalho apresenta as estratégias de detecção de pares de contatos duplicados e os algoritmos de agrupamento destes pares para futura integração de dados.

### **2 MATERIAIS E MÉTODOS**

Os registros são combinados em pares. Aqueles que compartilham pelo menos um número de telefone ou endereço de e-mail (casamento por igualdade) são identificados como duplicados. Sobre os campos dos demais registros são aplicadas as seguintes funções de similaridade (COHEN, RAVIKUMAR & FIENBERG, 2003): Levenshtein (login), Jaccard (nome), JaroWinkler (nome) e Monge-Elkan (nome). Os escores retornados são combinados por uma média ponderada. Se o valor resultante é maior que um determinado limiar de similaridade, os registros são considerados equivalentes, ou seja, representam contatos duplicados. Os pesos e o limiar são definidos como parâmetros de configuração.

### **3 RESULTADOS E DISCUSSÃO**

Foram propostas duas estratégias distintas para agrupar registros duplicados e, futuramente, integrar as informações em um único contato: (i) aglomerativa – cada registro é equivalente a pelo menos um registro do mesmo grupo e; (ii) restritiva – todos os registros de um grupo são equivalentes entre si. A figura 1 apresenta os algoritmos que implementam as duas estratégias descritas.

Considere o conjunto de registros de contatos representados pelos seus identificadores  $C = \{ 1, 2, \dots, 10 \}$  e os seguintes pares de registros detectados como duplicados  $PD = \{ [1,2], [3,4], [4,5], [5,6], [6,8], [7,8] \}$ . A estratégia aglomerativa resulta nos grupos de contatos equivalentes  $CE = \{ [1,2], [3,4,5,6,7,8], [9], [10] \}$ . Usando a estratégia restritiva são formados os grupos  $CE = \{ [1,2], [3,4], [5,6], [7,8], [9], [10] \}$ .

Figura 1 – Implementação das estratégias propostas

Algoritmo: Estratégia Aglomerativa

Entrada: conj. de contatos C, conj. de pares de contatos duplicados PD

Saída: conj. de grupos de contatos equivalentes GE

$GE = \emptyset$

$\forall [i,j] \in PD \forall G \in GE$

se  $i \in G$  então  $G = G \cup \{j\}$

senão se  $j \in G$  então  $G = G \cup \{i\}$

senão se  $\sim \exists G \in GE \mid i \in G \vee j \in G$  então  $GE = GE \cup \{[i,j]\}$

$\forall k \in C \forall G \in GE$

Se  $k \in G$  então  $C = C - \{k\}$

$GE = GE \cup C$

Retornar GE

Algoritmo: Estratégia Restritiva

Entrada: conj. de contatos C, conj. de pares de contatos duplicados PD

Saída: conj. de grupos de contatos equivalentes GE

$GE = \emptyset$

$\forall [i,j] \in PD \forall G \in GE$

se  $i \in G \wedge \forall e \in G \mid e \neq i, \exists [e,j] \in PD \mid e \neq j$  então  $G = G \cup \{j\}$

senão se  $\sim \exists G \in GE \mid i \in G$  então  $GE = GE \cup \{[i,j]\}$

$\forall k \in C \forall G \in GE$

Se  $k \in G$  então  $C = C - \{k\}$

$GE = GE \cup C$

Retornar GE

## 5 CONSIDERAÇÕES FINAIS

O algoritmo da estratégia restritiva resulta num número maior de grupos, mas é mais segura, porque só agrupa contatos similares entre si. Isto significa que as funções de similaridade devem ter classificado como duplicadas todas as combinações destes contatos dois a dois. Já o algoritmo da estratégia aglomerativa, apesar de gerar um número menor de grupos resultando numa agenda de contatos mais enxuta, pode inserir alguns registros erroneamente detectados como duplicados em determinados grupos de forma cumulativa, degradando a qualidade do agrupamento.

## REFERÊNCIAS

COHEN, W.; RAVIKUMAR, P.; FIENBERG, S. A comparison of string metrics for matching names and records. In: KDD Workshop on Data Cleaning and Object Consolidation, vol. 3, p. 73-78, 2003.

PINHEIRO, R.; LINDENAU, G.; ZIMMERMANN, A.; BORGES, E. N. Um aplicativo para integração de contatos em dispositivos Android. In: Anais do Congresso Regional de Iniciação Científica e Tecnológica em Engenharia, p. 1-4, 2014.