
Ain Shams University
Faculty of Engineering
Computer and Systems Engineering Department

# Cross-language Record Linkage for Big Data

A Thesis
submitted in partial fulfillment of the requirements of the degree of
Master of Science in Electrical Engineering

Submitted by
**Doaa Medhat Mohamed El-Saeed El-Mandouh**
B.Sc. of Electrical Engineering
(Computer and Systems Engineering Department)
Ain Shams University, 2009

Supervised by
**Dr. Ahmed Hassan Mohamed**
**Dr. Cherif Ramzi Salama**

**Cairo, 2016**

**AIN SHAMS UNIVERSITY**
**FACULTY OF ENGINEERING**
**Computer and Systems Engineering**

# Cross-language Record Linkage for Big Data

**by**

**Doaa Medhat Mohamed El-Saeed El-Mandouh**
**Bachelor of Science in Electrical Engineering**
**(Computer and Systems Engineering)**
**Faculty of Engineering, Ain Shams University, 2009**

**EXAMINERS' COMMITTEE**

| **Name and Affiliation** | **Signature** |
|---|---|
| **Prof. Mohamed Gamal El-Din Darwish**<br>**Faculty of Computer Science, Cairo University.** | ..... |
| **Prof. Hoda Korashi Mohamed Ismail**<br>**Computer and Systems Engineering Department**<br>**Faculty of Engineering, Ain Shams University.** | ..... |
| **Assoc. Prof. Ahmed Hassan Mohamed Yousef**<br>**Computer and Systems Engineering Department**<br>**Faculty of Engineering, Ain Shams University.** | ..... |

**Date:** / / 2016

# Statement

This dissertation is submitted to Ain Shams University for the degree of
Master of Science in Electrical Engineering (Computer and Systems Engineering).

The work included in this thesis was carried out by the author at the Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University, Cairo, Egypt.

No part of this thesis was submitted for a degree or a qualification at any other university or institution.

**Name:** Doaa Medhat Mohamed El-Saeed El-Mandouh

**Date:** / / 2016

# Researcher Data

**Name:** Doaa Medhat Mohamed El-Saeed El-Mandouh

**Date of Birth:** 23th of June 1986

**Place of Birth:** Cairo, Egypt

**First University Degree:** B.Sc. in Electrical Engineering

**Name of University:** Ain Shams University

**Date of Degree:** 2009

# Cross-language Record Linkage for Big Data

**Doaa Medhat Mohamed El-Saeed El-Mandouh**

Masters of Science Dissertation
Computer and Systems Engineering Department
Faculty of Engineering - Ain Shams University

## ABSTRACT

This thesis demonstrates the dire need for a powerful record linkage process to efficiently correlate data from different sources. It starts with an introduction about record linkage process with a survey on different techniques introduced in this area. It illustrates how the problem grows to be more complex when the goal is to manipulate big data. Subsequently, it presented the effectiveness and efficiency aspects. The former is needed for achieving high quality of matching records from different languages while the latter is needed for achieving a scalable load balanced record linkage process over large-scale multilingual data sources.

Afterword, the thesis introduces a novel technique relying on exiting pattern-based and phonetic matching techniques, which supports the matching of names written in different writing scripts effectively. Consequently, the thesis introduces a new cost-aware load balancing technique for achieving a better load balancing while matching large-scale multilingual data sources, which takes into consideration the different costs for matching cross-language records and mono-language ones. Finally, it applies the proposed techniques on some case studies, where they showed more effective and efficient results against existing techniques.

**Key words:** Record Linkage, Entity Matching, Cross Language, Multilingual, Big data, MapReduce.

**Faculty of Engineering – Ain Shams University**
**Computer and Systems Engineering Department**

Thesis title: "**Cross-language Record Linkage for Big Data**"

Submitted by: Doaa Medhat Mohamed El-Saeed El-Mandouh

Degree: Master of Science in Electrical Engineering

## Thesis Summary

This dissertation demonstrates the importance of Cross-language Record Linkage for Big Data. The dissertation is organized in eight chapters as follows:

**Chapter 1**
This chapter provides an overview about cross-language record linkage for Big Data. Motivation, objective and contributions of this work are presented. Also, the organization of the thesis is highlighted.

**Chapter 2**
This chapter illustrates a background about the record linkage process taking into account its different phases and the used techniques in each phase. In addition, it illustrates big data and distributed processing of large tasks that are required to scale up the record linkage process.

**Chapter 3**
This chapter discusses the state-of-art related work for cross-language matching, adaptation of the record linkage process to handle big data, and the existing approaches to load balance this process in the presence of workload imbalance

**Chapter 4**
This chapter introduces an overview about the proposed cross language record linkage process and the required adaptations to scale this process to work on Big Data.

**Chapter 5**
In this chapter, a generalized cross-language name matching framework is introduced which is used for matching names from different languages.

**Chapter 6**
In this chapter, an enhanced blocking-based record linkage process for Big Data is presented using a proposed cost-aware load balancing technique, which takes in consideration the different costs for matching mono-language and cross-language records in multilingual data sources.

**Chapter 7**
This chapter illustrates the different executed experiments and their results to evaluate the efficiency and effectiveness of the proposed solution. The experiments are performed on real data that is distributed on a cluster of multiple machines in a cloud environment.

**Chapter 8**
This chapter concludes the work presented in this thesis, and highlights the proposed future work.

# ACKNOWLEDGMENT

**All gratitude to ALLAH**

I would like first to thank my supervisors Dr. Ahmed Hassan and Dr. Cherif Salama for their insightful thoughts, continuous guidance, encouragement, help, and patience.

Many thanks to my colleagues and friends for their support and help during the work on this thesis.

Last but not least, I would like to thank all my family, especially my parents and sisters for supporting me through my whole life. Their encouragement, care, and love are what guided me to accomplish this work.

# Contents

# List of Figures