



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**

Faculty of Computer Science, Cairo University.

.....

**Prof. Hoda Korashi Mohamed Ismail**

Computer and Systems Engineering Department  
Faculty of Engineering, Ain Shams University.

.....

**Assoc. Prof. Ahmed Hassan Mohamed Yousef**

Computer and Systems Engineering Department  
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:** 23<sup>th</sup> 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

<b>List of Figures</b>	<b>VIII</b>
<b>List of Tables</b>	<b>X</b>
<b>List of Abbreviations</b>	<b>XI</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Scope and Contributions . . . . .	2
1.3 Thesis Organization . . . . .	3
<b>2 Background</b>	<b>5</b>
2.1 Record Linkage Process . . . . .	5
2.1.1 Data Pre-Processing Phase . . . . .	7
2.1.2 Blocking Phase . . . . .	8
2.1.3 Record-pair Comparison Phase . . . . .	10
2.1.3.1 Phonetic Matching Techniques . . . . .	11
2.1.3.2 Pattern Matching Techniques . . . . .	12
2.1.3.3 Dictionary-Based Matching Techniques . . . . .	14
2.1.4 Classification Phase . . . . .	15
2.1.5 Evaluation Measures . . . . .	15
2.2 Record Linkage for Big Data . . . . .	17
2.2.1 Big Data processing using MapReduce . . . . .	17
2.2.2 Blocking-based Record Linkage using MapReduce	20
<b>3 Related Work</b>	<b>22</b>
3.1 Cross-language matching . . . . .	22

3.2	Record linkage for Big Data . . . . .	25
3.2.1	Load balancing approaches for MapReduce based Record Linkage . . . . .	26
3.2.1.1	Detailed elaboration of BlockSplit and Multi-Source BlockSlicer Techniques . . . . .	30
<b>4</b>	<b>Overview of proposed Cross Language Record Linkage for Big Data Solution . . . . .</b>	<b>35</b>
4.1	Proposed Solution Overview . . . . .	35
4.1.1	Cross Language Matching . . . . .	35
4.1.2	Multilingual Record Linkage for Big Data . . . . .	36
<b>5</b>	<b>Proposed Cross Language Matching Solution . . . . .</b>	<b>38</b>
5.1	Introduction . . . . .	38
5.2	Language Identification . . . . .	40
5.3	Normalization . . . . .	40
5.3.1	General Normalization . . . . .	40
5.3.2	Language-Specific Normalization . . . . .	41
5.3.2.1	Arabic Characters Normalization . . . . .	41
5.3.2.2	Latin Characters Normalization . . . . .	41
5.4	Parsing . . . . .	43
5.5	Pair-wise comparison . . . . .	48
5.5.1	Phonetic Encoding . . . . .	48
5.5.2	Proposed Cross Language Levenshtein Distance . . . . .	50
5.5.3	Proposed additional features . . . . .	58
5.6	Classification . . . . .	59
5.7	Summary . . . . .	59
<b>6</b>	<b>Proposed Multilingual Record Linkage for Big Data Solution . . . . .</b>	<b>60</b>
6.1	Introduction . . . . .	60
6.2	Analysis MapReduce Job . . . . .	63
6.2.1	Map Phase (Blocking) . . . . .	65
6.2.2	Reduce Phase (Count Occurrences) . . . . .	68
6.3	Matching MapReduce Job . . . . .	71
6.3.1	Proposed Cost-Aware Load Balancing Mechanism . . . . .	71
6.3.1.1	Load Balancing Initialization . . . . .	72

6.3.1.2	Proposed Cost-Aware Block Splitting Mechanism . . . . .	73
6.3.1.3	Load Balancing of match tasks . . . . .	80
6.3.2	Map Phase (Proposed Cost-Aware Load Balancing)	83
6.3.2.1	Setup Execution . . . . .	83
6.3.2.2	Map Execution . . . . .	84
6.3.3	Reduce Phase (Proposed Multilingual Similarity Calculations) . . . . .	86
6.4	Evaluation MapReduce Job . . . . .	87
6.4.1	Map Phase (Source Tagging) . . . . .	89
6.4.2	Reduce Phase (Measure Calculations) . . . . .	89
6.5	Summary . . . . .	90
<b>7</b>	<b>Experimental Results and Discussion</b>	<b>91</b>
7.1	Environment Setup . . . . .	91
7.1.1	Data Preparation . . . . .	91
7.1.2	Distributed environment setup . . . . .	93
7.2	Experiments . . . . .	94
7.2.1	Experiments for Effectiveness Measurement . . .	94
7.2.1.1	Phonetic and Pattern Based Matching Comparison . . . . .	94
7.2.1.2	Hybrid Matching Technique: Components Analysis . . . . .	97
7.2.1.3	Summary . . . . .	99
7.2.2	Experiments for Efficiency Measurement . . . . .	100
7.2.2.1	Robustness to Data Skew . . . . .	101
7.2.2.2	Capability of load balancing Multilingual data sources . . . . .	107
7.2.2.3	Scalability to number of nodes . . . . .	111
7.2.2.4	Scalability to size of data . . . . .	114
7.2.2.5	Summary . . . . .	117
<b>8</b>	<b>Conclusion and Future Work</b>	<b>118</b>
8.1	Conclusion . . . . .	118
8.2	Future Work . . . . .	119

## *CONTENTS*

---

<b>Publications</b>	<b>120</b>
<b>References</b>	<b>121</b>
<b>Appendix</b>	<b>126</b>

---

# List of Figures

2.1	General Record Linkage Process . . . . .	7
2.2	Standard Blocking Example . . . . .	10
2.3	Levenshtein Distance Example . . . . .	14
2.4	MapReduce Process Illustration . . . . .	18
2.5	Blocking-based Record Linkage using MapReduce . . .	21
3.1	MapReduce-based linkage with load balancing [29] . . .	27
3.2	BlockSplit extension splitting example . . . . .	31
3.3	Multi-Source BlockSlicer splitting example . . . . .	32
3.4	Final workload distribution example . . . . .	33
5.1	Cross-Language Matching Solution Architecture . . . .	39
5.2	Parsing compound name with different forms . . . . .	44
5.3	Parse tree for full name with alternative name . . . . .	45
5.4	Parse tree for full name with different constituents . . . .	47
5.5	ArabicSoundex and Soundex Examples . . . . .	48
5.6	Proposed Extended ArabicSoundex and Soundex Examples	49
5.7	Proposed Algorithm for CLLD . . . . .	53
5.8	Example of Proposed CLLD . . . . .	54
6.1	Proposed Record Linkage Process Architecture . . . . .	61
6.2	Proposed Analysis MapReduce Job . . . . .	64
6.3	Match Task generation for small block . . . . .	73
6.4	Proposed Algorithm for Cost-Aware Block Splitting . . .	76
6.5	Proposed Cost-Aware Block Splitting Mechanism . . . .	77
6.6	Proposed Cost-Aware Block Splitting Mechanism - Match Tasks . . . . .	78
6.7	Greedy Optimization for Load Balancing the match tasks	81

## *LIST OF FIGURES*

---

6.8	Matching MapReduce Job . . . . .	82
6.9	Evaluation MapReduce Job . . . . .	87
6.10	Evaluation MapReduce Job Data flow . . . . .	88
7.1	Sample of generated JSON objects . . . . .	92
7.2	Quality measures for proposed hybrid techniques . . . .	97
7.3	Execution time for the proposed hybrid techniques . . .	98
7.4	Example of blocks distribution for different data skew . .	102
7.5	Robustness to Data Skew Experiment Results . . . . .	105
7.6	Capability of load balancing Multilingual data sources Experiment Results . . . . .	109
7.7	Scalability to number of nodes Experiment Results . . .	112
7.8	Scalability to size of data Experiment Results . . . . .	115