

DEVELOPMENT OF A COMPUTATIONAL SYSTEM FOR INTEGRATING USAGE INTO DOCUMENT INDEXING

By

LUKMAN ADEWALE AKANBI

M.Sc. (Computer Science), Ife



A THESIS SUBMITTED TO
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF TECHNOLOGY
OBAFEMI AWOLOWO UNIVERSITY, ILE-IFE

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD
OF DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

2014

OBAFEMI AWOLOWO UNIVERSITY ILE-IFE, NIGERIA

HEZEKIAH OLUWASANMI LIBRARY

POSTGRADUATE THESIS

AUTHORISATION TO COPY

Author: AKANBI Lukman Adewale

Title: DEVELOPMENT OF A COMPUTATIONAL SYSTEM FOR INTEGRATING
USAGE INTO DOCUMENT INDEXING

Degree: Ph.D. (Computer Science)

Year: 2014

I, AKANBI Lukman Adewale, hereby authorise the Hezekiah Oluwasanmi Library to copy my thesis in part or whole in response to request from individuals and or organisations for the purpose of private study or research.

Signature of Author and Date

CERTIFICATION

I, Lukman Adewale Akanbi with registration number TP08/09/R0033 in the Department of Computer Science and Engineering, Faculty of Technology, Obafemi Awolowo University certify that this is an original research carried out by me under the supervision of:

Supervisor:

Prof. E. R. Adagunodo

Co-Supervisor:

Prof. A. A. David

Head of Department:

Dr. A. I. Oluwaranti

DEDICATION

This work is dedicated to glory of Almighty Allah, by whose permission the work have been completed.

OBAFEMI AWOLOWO UNIVERSITY

ACKNOWLEDGEMENT

All glorification and adoration are due to Allah, the Almighty, who out of His mercy, made the dream of having a Doctorate degree became reality. I am most grateful to my employer, the Obafemi Awolowo University, Ile-Ife for giving me the opportunity and the required supports. My sincere appreciation goes to the Heads of Department of Computer Science and Engineering, past and present for their support and advice throughout the course of the research work. I am particularly grateful to all the members of staff of the Department, for their supports and holding forth for me when I had to travel.

My profound gratitude goes to my supervisors, Professors Emmanuel Rotimi Adagunodo and Amos Abayomi David, for their time and resources. I appreciate your fatherly role and the sacrifices you have to make throughout the course of the work. I pray God almighty reward you with good in manifolds. The effort of Prof. David at securing grant that made it possible for me to visit France twice during the course of the work is highly appreciated. His hosting me in France and making my staying in France very rewarding is highly appreciated.

My appreciation goes to the French Government through the French embassy in Nigeria, for supporting this work through the provision of grant that facilitated my visit to France in 2012 and 2013. I also appreciate my colleagues in France, Toyin Oguntunde and Bunmi Akere for making me feeling at home in France.

I am most grateful to Dr. B. S. Afolabi for his immense contribution towards the success of this work. The advice and inspiration from Dr. O. A. Odejebi is highly appreciated. I am also grateful to Drs. S. I. Eludiora, A. O. Ajayi, F. O. Asahiah, S. A. Bello, Mrs. K. C. Olufokunbi, Mrs. D. F. Ninan, Mr. J. A. Hassan and other members of the CISRG for their supports during the

course of the research work. Dr. AdulWaheed Bamgbade is highly appreciated for his professional support through timely review of the grammatical structure of the thesis.

My friends, AbdulAzeem Ewenla, Lukman Olawoyin, Azeez Adebisi, Taofeek Olusesi, Mikail Farinde and others are appreciated for their love and care. I thank you all. The contribution of Mahmood Oyewo in coding at the initial stage of the work is highly appreciated.

I am particularly grateful to my loving wife, Mrs. Adebimpe Monsurah Akanbi and my beautiful daughter Rodiat Adeola Akanbi for their patient and support throughout the course of this research work. I cannot appreciate you enough. I love you. My appreciation goes to other members of my family for their understanding and support in the course of this study.

TABLE OF CONTENTS

Title Page.....	i
Authorisation to Copy.....	ii
Certification.....	iii
Dedication.....	iv
Acknowledgement.....	v
Table of Contents.....	vii
List of Tables.....	ix
List of Figures.....	x
Abstract.....	xvi
CHAPTER ONE: INTRODUCTION.....	1
1.1 Background to the Study.....	1
1.2 Statement of Research Problem.....	8
1.3 Research Aim and Objectives	9
1.4 Research Methodology	9
1.5 Research Philosophy.....	10
1.6 Research Motivation	10
1.7 Scope of Research.....	11
1.8 Organisation of Thesis	11
CHAPTER TWO: LITERATURE REVIEW.....	12
2.1 The Concept and Definition of Document	12
2.1.1 Types of Document	16
2.1.2 Classification of Document	16

2.2 Competitive Intelligence	18
2.2.1 Actors in Competitive Intelligence Process	20
2.2.2 Competitive Intelligence System Architecture	22
2.3 Information Retrieval.....	26
2.3.1 Boolean Model of IR	26
2.3.2 Vector Model of IR	28
2.3.3 Probabilistic Model of IR	30
2.4 Document Indexing.....	30
2.5 Document Annotation	31
2.6 Document Annotation Tools	32
2.6.1 ComMentor annotation tool.....	32
2.6.2 Yawas annotation tool	35
2.6.3 Annotea annotation tool.....	36
2.6.4 AMIE annotation tool.....	40
2.6.5 AMTEA annotation tool.....	45
2.7 The Fuzzy Logic	46
2.7.1 Fuzzy set	46
2.7.2 Linguistic variable.....	49
2.7.3 Membership function.....	49
2.7.4 Operations with fuzzy sets	52

2.7.5 Fuzzy rules	55
2.7.6 Structure of fuzzy inference system	55
2.7.7 Procedure of fuzzy reasoning.....	58
2.8 Related Works on Document Indexing	58
2.9 Chapter Summary	62
CHAPTER THREE: MODEL DEVELOPMENT.....	63
3.1 Overview	63
3.2 Architecture of the CIDUCE System.....	63
3.2.1 The Information World.....	65
3.2.2 The Information Base	65
3.3 Structure of the Proposed System.....	66
3.3.1 Usage Creation Module	68
3.3.2 Usage Exploration Module	71
3.4 Document Usage Model.....	73
3.4.1 The User (<i>U</i>)	74
3.4.2 Decision Problem (<i>P</i>)	75
3.4.3 The Document (<i>D</i>)	77
3.4.4 The Environment (<i>E</i>).....	78
3.5 Document Degree of Relevance to the Resolution of DPs	79
3.5.1 Fuzzification Process for the Input Data	79

3.5.2 The Output Data	86
3.5.3 Data for the Fuzzy Logic Model	89
3.5.4 Fuzzy Inference Engine	91
3.5.5 Defuzzification Process	96
3.6 The Document Usage Index	97
3.7 Document Usage Model Evaluation	98
3.7.1 Data for the Usage Model Evaluation	102
3.7.2 Data Extracted from the Questionnaire	102
3.8 Chapter Summary	104
CHAPTER FOUR: SYSTEM DEVELOPMENT AND EVALUATION	105
4.1 Overview	105
4.2 Object-Oriented Model of the CIDUCE System.....	105
4.2.1 Use Case Diagram.....	105
4.2.2 Class Diagrams	106
4.2.3 Sequence Diagrams	112
4.3 System Prototype Implementation.....	115
4.3.1. Hardware Requirements	119
4.3.2 Software Requirements.....	120
4.4 User Interface of CIDUCE System.....	120
4.5 Model Evaluation and Result Discussion	127

CHAPTER FIVE: SUMMARY, CONCLUSIONS AND RECOMMENDATIONS	145
5.1 Summary	145
5.2 Conclusion.....	146
5.3 Contribution to Knowledge	146
5.4 Application Area and Future Work	147
REFERENCES	149
APPENDIX A: QUESTIONNAIRE.....	158
APPENDIX B: DATA EXTRACTED FROM THE QUESTIONNAIRE	168
APPENDIX C: DOCUMENTS AND DECISION PROBLEMS REPRESENTATION IN VSM FOR RESPONDENTS	180
APPENDIX D: RESULTS OF SIMILARITY MEASURE BETWEEN DECISION PROBLEMS AND DOCUMENTS FOR OTHER RESPONDENTS	221

LIST OF TABLES

1.1: Information Search through Web Search Facilities	3
2.1: Categorization of Documents	18
4.1: Document Upload Use Case Scenario	113
4.2: Scenario for Usage Creation Use Case	114
4.3: Scenario for Cross Analysis Use Case	115
4.4: Document by Term Matrix in VSM for keyterm based index for Respondent-1	137
4.5: Document by Term Matrix in VSM for usage based index for Respondent-1	138
4.6: Result of Similarity Measure between DP and Documents for Respondent-1	140
4.7: Result of Similarity Measure between Thesis Title and Documents for Respondent-1	143

LIST OF FIGURES

1.1: A Sample Information Search through Web Search Facilities (Google) with plain query terms	4
1.2: A Sample Information Search through Web Search Facilities (Google) with query term quoted	5
2.1: Document Use at Different Stage of Research Project.....	16
2.2 Relationship between Decision Maker and Information Watcher.....	24
2.3: Competitive Intelligence System Architecture.....	25
2.4: The Architecture of ComMentor Annotation Tool.....	35
2.5: Screen Shot of Yawas for Creating New Annotation.....	38
2.6: Using the Context Menu to call the Yawas Options.....	39
2.7: The Basic Architecture of Annotea.....	42
2.8: Information Search from the Annotation Extension.....	44
2.9: Logical view of Annotation-as-a-process.....	48
2.10: An Example for membership function - <i>positive small temperature</i>	52
2.11: Universe of Discourse for Linguistic Variable temperature.....	53
2.12: Membership functions represented with triangular shape (a) with minimum operator and (b) with maximum operator.....	56
2.13: Structure of Fuzzy Inference System.....	58

2.14: Structure of the FUZZY Part of the System.....	59
3.1: Architecture of the CI-DUCE System	68
3.2: Structural Model of the CI-DUCE System	70
3.3: Flowchart of the CI-DUCE System	72
3.4: Flowchart of Document Usage Creation Module	74
3.5: Flowchart of Document Usage Exploration Module	76
3.6: Fuzzy Logic for Deriving Document Relevance to DP	85
3.7: Membership Function for NSP handled by the user	86
3.8: Membership Function for Number of Years Spent.....	89
3.9: Membership Function for Users' Specified Degree of Relevance of Document to DP	91
3.10: Membership Function for Document Degree of Relevance.....	94
3.11: The Rule Base of the FL Inference Engine in Tabular form.....	97
3.12: The Rule Base of the FL Inference Engine.....	98
3.13: Formal description of the usage-based document representation schemes.....	104
3.14: Sample Data Extracted from the Questionnaire.....	108
4.1: Use Case Diagram of the CIDUCE System.....	112
4.2: System Class Diagram showing Composition Relationship.....	116
4.3: System Class Diagram showing Association.....	118
4.4: Sequence Diagram for Search Operation.....	119
4.5: Sequence Diagram for Usage Based Search Operation.....	121
4.6: Sequence Diagram for Usage Creation Operation.....	122
4.7: Sequence Diagram for Usage Exploration Operation	123
4.8: Home Page of the CIDUCE System.....	126

4.9: Interface for Adding Usage to a Document.....	128
4.10: Interface for Cross Analysis Task.....	129
4.11: Result of Cross Analysis Task.....	130
4.12: Graphical Representation of Cross Analysis Result.....	131
4.13: Interface for Search Operation	133
4.14: Result of Normal Search Operation	134
4.15: Result of Usage Based Search Operation.....	135
4.16: Similarity between DP and Documents.....	142
4.17: Number of Relevant Documents to DP at Different Thresholds.....	145
4.18: Number of Relevant Documents to DP at Different Thresholds in the case of Thesis Titles as DPs	146

ABSTRACT

The study formulated a model that augments document with usage, designed, implemented and evaluated a system based on the model. This is with the view of enhancing the quality and quantity of useful documents that are returned during document search operation.

Attribute Value Pair technique of data abstraction in document annotation and vector model technique of Information Retrieval were used to formulate the document usage model. Unifying Modelling Language (UML 2.0) was used to design the Competitive Intelligence based Document Usage Creation and Exploration (CIDUCE) system. The prototype was implemented with the use of PHP and MySQL technology. Data on document usage was collected through questionnaire administration and guided interview from 20 selected postgraduate students (M.Sc. and Ph.D.) in various departments in the Faculty of Technology. Ninety-nine (99) documents and twenty (20) decision problems were extracted from the questionnaire and used to populate the database of the system. Document recall rate, a function of the similarity measure between identified relevant documents by the respondents and their decision problems (i.e. research problems) was used to evaluate the system.

The results showed that the usage-based document index consistently produce high recall rate, that is, identified high number of relevant documents at different retrieval thresholds than the keyterm-based index. For example, at the retrieval thresholds of 0.20, 0.30, 0.40, 0.50, 0.60, 0.70 and 0.80, the keyterm-based index has 47.47, 27.27, 14.14, 9.09, 2.02, 1.01 and 0.00% recall rates, respectively as compare with the usage-based index with recall rate of 100.00, 100.00, 100.00, 100.00, 100.00, 91.92 and 61.62%, respectively. These recall rates at different thresholds translated to 47, 27, 14, 9, 2, 1 and 0 documents, respectively in the keyterm-based index and 99, 99, 99, 99, 92 and 62 documents, respectively in the usage-based index.

The study concluded that in an information seeking process, there are usually documents in the document collection space whose index may not contain terms in the users query but which are very relevant to users' need.

OBAFEMI AWOLOWO UNIVERSITY