

Acknowledgements

I would firstly like to thank my supervisor Awad EL-Kareem Yousoffor his wise advice and foresight in letting me run with my own ideas and allowing them to flourish. I must also thank Ahmed Saad for his support of my work and thoughtful ideas.

Dedication

"Who does not thank people does not thank God" believing in this, I readily dedicate this research to my mother, the always flowing source of liberality, and to the spirit of my late beloved father, may Allah rest his soul in eternal peace and clammy mercy, and to all who stood by me in my procession my wife and my children loved Saad, Mohammed and daughter Rayan. I can never forget my instructors and supervisors on this research and my utmost thanks and praises to God Almighty.

Abstract

Efficient techniques to detect similar data in many data sources has become one of the most important and challenging issues in many areas such as Database, Bioinformatics and Data Mining. In this research, a three phase framework for similarity detection is proposed:

In the first phase:

Data Sources were collected from the web, depending on how it relates to a predetermined domain. The base source is the source of the data available, which describes the domain.

In the second phase:

the sources obtained are filtered to select data sources with a greater probability of containing data describing the domain by examining the degree of similarity between the base source, and each source from the sources obtained "External Sources". Whereas the selection is only for the external sources which its `simi_degree` value is less than, or equal to the average of the `simi_degree` values of all sources.

In the third phase:

Content similarity is examined between the base source, and all the selected external sources in phase 1, by using the proposed "Probability Measure" that gives a value on the basis of which it is determined whether the content of external sources is similar to the content of the base resource.

Experimental result shows that the researcher's similarity framework can achieve better quality result than the conventional approaches.

مستخلص البحث

إيجاد طريقة فعالة لفحص التشابه بين مصادر البيانات المتعددة من القضايا المهمة التي تمثل تحدياً يواجهه عدة مجالات مثل قواعد البيانات ونظم المعلومات الحيوية ومجال التنقيب عن البيانات.

في هذا البحث يقدم الباحث إطار عمل يتكون من ثلاثة مراحل لفحص التشابه بين مصادر البيانات.

المرحلة الأولى

تجميع مصادر البيانات اعتماداً على علاقتها مع المجال المحدد سلفاً. المصدر الأساسي هو مصدر البيانات الذي يصف المجال ويحتوى على بيانات نبحث عن ما يشابهها في المصادر الأخرى.

المرحلة الثانية

تصفية مصادر البيانات التي تم الحصول عليها في المرحلة الأولى وذلك بحساب درجة التشابه بين مصادر البيانات الخارجية والمصدر الأساسي كل على حدا. حيث يتم اختيار المصادر التي تكون قيمة درجة التشابه لها أقل من أو مساوية لمتوسط قيم درجات التشابه ككل.

المرحلة الثالثة

تتعلق هذه المرحلة بفحص محتوى المصادر التي تم الحصول عليها في المرحلة الثانية باستخدام مقياس إحصائي هذا المقياس يعطى قيمة على أساسها يتم تحديد ما إذا كان المحتوى مشابه أم لا، ونحصل على التشابه عندما تكون القيمة التي ينتجها المقياس أكبر من واحد. اثبتت النتائج التي حصل عليها الباحث من خلال التطبيق إن الإطار المقترح يعطى نتائج أفضل من الطرق التقليدية المستخدمة.

Table of Contents

Subject	page No.
Acknowledgements	I
Dedication	II
Abstract	III
Arabic Abstract	IV
Table of contents	V
List of tables.....	IX
List of figures	X

CHAPTER 1: INTRODUCTION

1.1 Problem overview	1
1.2 Research Problem	2
1.3 Objectives	4
1.4 Research Scope	5
1.5 Contributions	5
1.6 Organization of Thesis	6

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction	7
2.2 Concept in philosophy and Psychology	8
3.3 Similarity and Mathematics	9
2.4 Areas of similarity Detection.....	9
2.5 Distance-Based Similarity Measures	9
2.5.1 Euclidean Distance.....	10
2.5.2 Pearson's Correlation Coefficient.....	10
2.5.3 Jaccard Similarity Coefficient.....	10
2.5.4 Tanimoto Coefficient	11

2.5.5 Cosine Similarity.....	11
2.5.6 Dice’s Coefficient.....	12
2.6 Feature-Based Similarity Measures	12
2.6.1 Contrast Model	12
2.7 Probabilistic Similarity Measures	13
2.8 The Principles and Practice of Numerical Classification.....	14
2.9 Similarity Measures for Categorical Data.....	15
2.10 Mining Heterogeneous Transformations for Record Linkage	15
2.11 An Object Oriented Email Clustering Model Using Weighted Similarities between Emails Attributes	16
2.12 Weight Similarity Measurement Model Based, Object Oriented Approach for Bug Databases Mining to Detect Similar and Duplicate Bugs.....	16
2.13 Measuring Similarity Between Collection of Values.....	17
2.14 DogmatiX.....	18
2.15 Improving Semi-Supervised Acquisition of Relation Extraction Patterns.....	19
2.16 Searching multiple databases for interesting complexes..	20
2.17 Information source selection methodology for recommender systems	20
2.18 learning from multiple sources	20
2.19 Matching and integration across heterogeneous data source	21
2.20 Learning model for multi-source integration.....	21
2.21 Sequential Pattern Mining in Multi-Databases via Multiple Alignment	22
2.22 Querying Multiple Bioinformatics Information Sources.....	22
2.23 Similarity Measures for Multi-Valued Attributes for Database Clustering.....	22
2.24 Data Integration for Many Data Sources using Context- Sensitive Similarity.....	23
2.25 A new Similarity Measure for Instance Data Matching.....	23
2.26 A novel Similarity Measure for Clustering Categorical Set.	24

CHAPTER 3: METHODOLOGY

3.1 Introduction.....	26
3.2 Research Process and Methods.....	26
3.2.1 Source Selection	26
3.2.2 Content similarity	27
3.2.3 Testing	27
3.2.4 Evaluation	27
3.3 Research Assumptions.....	30

CHAPTER 4: PROPOSE MODEL

4.1 Introduction	31
4.2 Source Selection.....	31
4.2.1 Xml and sources modeling	33
4.2.2 Evaluating the selection method	38
4.3 Measuring content similarity	39

CHAPTER 5:IMPLEMENTATION

5.1 Source Selection	46
5.2 Source Selection Against Database Selection.....	46
5.3 The Base Source	46
5.4 Refining external sources.....	50
5.5 Evaluating selection process	54
5.6 Categorical attributes and Non- Categorical attributes.....	57
5.6.1 Categorical Attributes.....	57
5.6.2 Non- Categorical Attributes.....	58
5.7 Measuring the content of data sources.....	60
5.8 XML Reaper.....	69
5.8.1 Features	70
5.8.2 Inputs.....	70

5.8.3 Operation.....	70
5.8.4 Outputs.....	71
5.8.5 Concerns.....	71
5.9 Experiment.....	71
5.9.1 Experimental Data	73
5.9.2 Source selection	74
5.9.3 Content similarity detection.....	76
CHAPTER 6: CONCLUSION AND FUTURE WORK	
6.1 Conclusion.....	87
6.2 Future work	88
REFERENCES	
References	89

List of Table

Table 5.1	Hepatitis diagnostic data	47
Table 5.2	Website where data sources obtained	49
Table 5.3	Result of comparison process	52
Table 5.4	cos_sim and simi_degree values for external sources	56
Table 5.5	A set of s_value obtained by the XMLReaper	69
Table 5.6	values of simi_degree, Cosine_value and the number of similar features.....	74
Table 5.7	s_values and source of similarity values between (D ₃ and base source).....	78
Table 5.8	s_values and source of similarity values between (D ₄ and base source).....	79
Table 5.9	s_values and source of similarity values between (D ₅ and base source).....	80
Table 5.10	s_values and source of similarity values between (D ₆ and base source)	81
Table 5.11	s_values and source of similarity values between (D ₇ and base source)	82
Table 5.12	s_values and source of similarity values between (D ₈ and base source)	83
Table 5.13	s_values and source of similarity values between (D ₉ and base source)	84
Table 5.14	s_values and source of similarity values between (D ₁₀ and base source)	85
Table 5.15	XML Reaper result	86

List of Figures

Figure 3.1	The propose model.....	28
Figure 3.2	The Research Process	29
Figure 4.1	Transformation Process.....	35
Figure 4.2	Similarity analyzer	40
Figure 4.3	Steps of finding similar sources to extract data.....	43
Figure 5.1	Source domain ant its sub_domains	60
Figure 5.2	Relation between simi_degree value and the number of similar features between base source and external sources.....	75
Figure 5.3	relation between cosine values and the number of similar features between base source and external sources	76