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 يواجه 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 الأولي

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 الاخرى.

المرحلة الثانية

تصفية مصادر البيانات التي تم الحصول عليها في المرحلة الأولي 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 بإستخدام 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 افضل من الطرق التقليدية المستخدمة.

Table of Contents

CHAPTER 3: METHODOLOGY

CHAPTER 4: PROPOSE MODEL

CHAPTER 5:IMPLEMENTATION

List of Table

List of Figures