جامعة الموصل

كلية علوم الحاسوب والرياضيات

استخلاص العلاقات النحوية من الويب باستخدام الحوسبة المرنة

أطروحة تقدمت بها

غادة عبد الكريم عبد العزيز

إلى

مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل

وهي جزء من متطلبات نيل شهادة دكتوراه فلسفة

في علوم الحاسوب

بإشراف

الأستاذ المساعد الدكتور

غيداء عبد العزيز الطالب

الأستاذ المساعد الدكتور

الاء ياسين طاقة

1441 هـ

2020م

الملخص

Abstract

إنَّ ازدياد استخدام الويب من قبل فئات مختلفة من المجتمع ودخوله في الحياة اليومية وبشتى المجالات أدى الى ظهور العديد من التحديات ، الأمر الذي جعل محركات البحث بأنواعها من أهم مصادر المعلومات ، وذلك عن طريق تصفح العدد الهائل من صفحات الويب بأشكالها وانواعها كافة. وقد ساعدت معالجة اللغات الطبيعية برمجيا فئات المجتمع غير المتخصص التفاعل مع بعضها البعض والاستفادة من خزائن المعلومات المتاحة على الويب.

يهدف البحث الى تصميم نظام لاسترجاع ملفات من الويب تتعلق بمجال الموسيقى وذلك بناء على استعلام المستخدم . تم اعتماد خوارزميات معالجة اللغات الطبيعية Natural Language Processing (NLP) وباستخدام لغة البايثون في استخلاص العلاقات النحوية الثلاثية Subject-Verb-Object (SVO) واستخلاص العلاقات الثنائية Subject-Verb (SV) و Verb-Object (VO)من نصوص ملفات الويب ذات الامتدادات (PDF , txt , html , DOC) وتم اعتماد هذه العلاقات النحوية كمفتاح اساسي في عملية استرجاع الملفات المطلوبة .

استخدمت خوارزمية مصنف بيز البسيط لتحديد الصنف الذي ينتمي اليه الاستعلام فيما اذا كان ذا علاقة بالموسيقى او لا ومن ثم استرجاع الملفات المرتبطة بموضوع الاستعلام وحسب المسافة الاقليدية بين العلاقة الناشئة من الاستعلام والعلاقات في قاعدة المعرفة. درب المصنف على مجموعتين من العلاقات استخلصت من (1000) ملف تم انتقاءهم بشكل عشوائي من ملفات الويب. وكان عدد العلاقات المستخلصة لصنف الموسيقى يساوي (98500) علاقة ثلاثية و (197000) علاقة ثنائية في حين كان عدد العلاقات المستخلصة من الملفات التي ليس لها علاقة بالموسيقى تساوي (87500) علاقة ثلاثية و (175000) علاقة ثنائية. بلغت نسبة الدقة (precision) في بعض حالات الاختبار الى 98% ونسبة الاسترجاع (recall) هي (96%) ونسبة مقياس F1 (96%) وترجع هذه النسبة العالية للدقة والاسترجاع الى ان البحث عالج واحدة من اهم مشاكل استرجاع المعلومات (Information Retrieval) وهي الاستعلامات الناقصة والضعيفة وذلك باعتماد العلاقات الثنائية وليس الثلاثية فقط كاساس للاسترجاع.

UNIVERSITY OF MOSUL
COLLEGE OF COMPUTER SCIENCES
AND MATHEMATICS

Web Syntax Relations Extraction Using Soft Computing

Ghada Abdul Karim Abdul Aziz

Ph.D. / Thesis

Computer Science

Supervised by

Dr. Ghayda Abdul Aziz Altalib
Assistant Professor

Dr. Alaa Yaseen Taqa
Assistant Professor

2020 A.D. **1441 A. H.**

Abstract

The increased use of Web by different groups of society and its pervasiveness in various fields of everyday life have led to the emergence of many challenges. This, in turns, makes many of search engines the most important sources of information through browsing massive and diverse web pages. In addition, processing natural languages by software has assisted even in experts (ordinary people) to interact with each other and benefit of the wealth of information available on the internet.

The aim of this study was to design a retrieval system for music-related files based on the user's query. Natural Language Processing (NLP) algorithms were adopted using Python language compiler to extract the ternary relation Subject-Verb-Object (SVO) and binary relations Subject-Verb (SV) and Verb-Object (VO) from the text in files of the format (PDF, txt, html, DOC). These relations were used as key for retrieving required files.

The naive Bayes classifier algorithm was used to determine the category of the query and whether it was related to music. Then, files related to the query were retrieved according to the Euclidean distance between the relationship arising from the query and the relationships in the knowledge (data) base.

The work was conducted on two sets of relationships extracted from 1,000 randomly selected web files.

The number of relationships derived for the musical files was equal to (98,500) ternary relations and (197,000) binary relations, while the number of relationships derived from non-musical files was equal to (87,500) ternary relations and (175,000) binary relations. The precision rate in some cases reached 98% while the recall rate is 96% and the F1 measure is equal to 96%. This high rate of retrieval can be attributed to the fact that the study addressed one of the most important problems of queries, namely, incomplete or weak queries by adopting binary relations as a basis for retrieval as well as using the ternary relations alone.