وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم علوم الحاسوب

# استخراج الأحداث من مواقع الويب: تطبيق على موقع كلية علوم الحاسوب والرياضيات في جامعة الموصل

رسالة مقدمة
إلى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة ماجستير علوم في
علوم الحاسوب

من قبل
رشا وائـل علي النحاس

بإشراف
م. د. بان شريف مصطفى

١٤٤٤ هـ ٢٠٢٣م

## الملخص

مع انتشار استخدام خدمة الانترنت، واتساع المناطق التي تغطيها هذه الخدمة، أصبح الاعتماد عليها اعتماداً كبيراً في نشر الأحداث على مواقع الويب، وتختلف أنواع هذه المواقع وفق نشاطات المنظمة أو المؤسسة الراعية أو المالكة لهذه المواقع، التي تنشر الأحداث المهمة حول نشاط المنظمة؛ ليطلع المستفيدون على محتواها.

وزادت أعداد المواقع التي تنشر الاحداث المهمة المرتبطة بالمؤسسات المختلفة، وهذه الزيادة تحتاج إلى تقنية ضرورية تساعد على تحديد الحدث في موقع الويب، وتحديد محتوياته، لغرض تكوين قاعدة معرفية يمكن الاعتماد عليها في مقارنة النشاطات بين المؤسسات، والتعرف على نشاطاتها بشكل مختصر يؤدي الغرض المطلوب، وقدمت العديد من الدراسات التي عُنِيتْ بمعالجة اللغة الطبيعية التي كتبت بها الأحداث في مواقع الويب، وتطورت هذه التقنيات تطوراً متسارعاً أدى إلى ظهور طرائق عديدة لاستخراج الحدث، تعتمد على مهام المعالجة الطبيعية والنماذج المدربة لتحديد محفزات الحدث وتوابعه والعلاقة بينها، والتعرف على أسماء الكيانات التي يتكون منها النص في موقع الويب.

وكلية علوم الحاسوب والرياضيات في جامعة الموصل، من المؤسسات الاكاديمية الرصينة التي تنشر نشاطات الدراسات العليا على موقع الكلية، إضافة الى الحلقات النقاشية العلمية التي أقامتها الكلية، وبالنظر لكثر هذه النشاطات في أقسام الكلية المختلفة، أصبح من الضروري أن تستخدم تقنيات التعلم العميق ومعالجة اللغة الطبيعية في استخراج الحدث من موقع كلية علوم الحاسوب والرياضيات.

في هذه الرسالة، عالج النظام المقترح الأحداث المنشورة على موقع الكلية، باستخدام مهام متعددة لمعالجة اللغة الطبيعية، بدءًا من المعالجة الأولية، التي تضم قراءة صفحة من موقع الويب، وتحديد النص المتعلق بالحدث، وإزالة الفراغات والمقاطع البرمجية التي لا علاقة لها بمحتوى الحدث، ثم إزالة علامات التنقيط من النص، وتنتهي هذه المرحلة بتقطيع النص إلى مجموعة من الجمل.

وقد حددت الرسالة المحفزات الرئيسة والفرعية التي تشير إلى الحدث المطلوب إيجاده في موقع الويب، واستعملت هذه المحفزات في اكتشاف الحدث وتوابعه المرتبطة به للحصول على المعلومات المطلوبة، بالاعتماد على نماذج مدربة مسبقاً باللغة الإنكليزية، وهي تسمية الدور الدلالي(SRL) Semantic Role Labeling وتمييز الكيانات المسماة Named Entity Recognition (NER)، وهما من النماذج الفعالة في تحديد الحدث وتحديد محتوياته.

لقد صادف النظام بعض التحديات عند تطبيقه تتمثل بالأخطاء الإملائية والقواعدية، وتباين الأساليب الإنشائية في كتابة نص الحدث، لذلك تم تصميم مجموعة من الدوال الإجبارية التي تعتمد على المحفزات والتوابع في إيجاد المعلومات المفقودة خلال تطبيق النظام في مواقع فيها بعض هذه المشكلات.

وبينت نتائج التطبيق أن النظام المقترح حقق مستوى أداء جيد، فكانت مستويات الدقة تتراوح ما بين (٧٥ – ١٠٠%)، وهي في أقل حالتها نسب مفيدة، يمكن تطوير النظام لزيادتها، واقترحت الرسالة مقترحات علمية يمكن أن تثري مجال استخراج الحدث.

Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and Mathematics
Department of Computer Science

# Extract Events from Web Sites: Applied on Computer Science and Mathematics in University of Mosul

A Thesis Submitted to the Council of the College of
Computer Science and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Master of Science
in
Computer Science

By
**Rasha Wael Ali Al-Nahhas**

Supervised by
**Dr. Ban Shareef Mustafa**

2023 A.D. 1444 A.H.

## Abstract

With the widespread use of the Internet service, and the expansion of the areas covered by this service, reliance on it has become highly dependent in publishing events on websites, and the types of these sites differ according to the activities of the organization or institution that sponsors or owns these sites, which publish important events about the activity of the organization; so that the beneficiaries can see its content.

The number of websites that publish important events related to different institutions has increased, and this increase requires a necessary technology that helps identify the event in the website and its contents, for the purpose of creating a knowledge base that can be relied upon in comparing activities between institutions, and identifying their activities in a brief manner that serves the required purpose. Several studies were presented dealing with natural language processing in which events were written on websites, and these techniques developed rapidly, leading to the emergence of many methods for extracting the event, relying on natural processing tasks and trained models to identify event triggers and its dependencies and the relationship between them, and to identify the names of the entities that make up the event. Including the text in the website.

The College of Computer Science and Mathematics at the University of Mosul is one of the solid academic institutions that publishes postgraduate activities on the college's website, in addition to the scientific discussion panels held by the college. Given the large number of these activities in the various departments of the college, it has become necessary to use deep learning techniques and natural language processing. In extracting the event from the website of the College of Computer Science and Mathematics.

In this thesis, the proposed system processed the events published on the college website, by performing various natural language processing tasks, starting with the initial processing, which includes reading the event page from the website, identifying the text related to the event, removing spaces and code segments that are not related to the content of the event, then Removing punctuation marks from the text, and this stage ends with cutting the text into a group of sentences.

The thesis identified the main and sub triggers that refer to the event to be found on the website, and these triggers were used to discover the event and its related dependencies to obtain the required information, based on pre-trained models in English, namely Semantic Role Labeling (SRL) and entity recognition. Named Entity Recognition (NER), and they are effective models in identifying the event and determining its contents.

The system has encountered some obstacles in its application, represented by spelling and grammatical errors, and showing constructive methods in writing the text of the event. Therefore, a set of compulsory functions has been designed that rely on incentives and functions to find missing information during the application of the system in locations where some of these problems exist.

The results of the application showed that the proposed system achieved a good level of performance, so the levels of accuracy ranged between (75-100%), which are in their lowest case useful percentages, the system can be developed to increase them, and the thesis suggested scientific recommendations that could enrich the field of event extraction.