UNIVERSITY OF SCIENCE AND TECHNOLOGY

COLLEGE OF GRADUATE STUDIES AND ACADEMIC ADVANCEMENT

Faculty of Computer Science and Information Technology

Applying Clustering and Classification in Web Usage Mining (Case Study: British Government Website)

by

Osama Mohamed Ali Abu Elhassan

A Thesis

Submitted to the College of Graduate Studies and Academic Advancements in Partial Fulfillment of the Requirement for the Degree of Master of Computer Science

Supervisor

Dr. Atif Ali Mohamed

February 2016

Abstract

Nowadays, the Web has become one of the most widespread platforms for information change and retrieval. As it becomes easier to publish website , as the number of users, and thus publishers, increases and as the number of websites grows, extract for information is turning into a cumbersome and time-consuming operation. Due to heterogeneity and unstructured nature of the data available on the WWW, Web mining uses various data mining techniques and tools to discover useful knowledge from Web hyperlinks, page content and usage log.

Web Usage Mining is the process of applying data mining techniques to the discovery of usage patterns from Web data. Like other data mining disciplines, it defines several procedures leading to the discovery of the desired knowledge.

The aim of this research is to predict the user's interaction with the web site (data.gov.uk). The methodology used in this research is as follow *first* dataset is downloaded from the United Kingdom government site.

Second apply clustering technique to clustering this data into three groups and assume this groups is (high-interaction, low interaction and medium interaction) after labeled this data we used classification technique to make a rule to classification new data (testing data) and predicted the changes that may occur on the resource downloads category. (An increase or decrease in the size of one of the groups).

The model is applied using naïve base algorithm and the result explains: Cluster 0 (high-interaction category) is 283 datasets. Cluster 1 (low interaction category) is 3 datasets. Cluster 2 (medium interaction category) is 14 datasets, Can take advantage of this information as a service to customers or be useful information to competitors in the same field.

المستخلص

في الوقت الحاضر أصبح الويب واحد من المنصات الاكثر انتشارا لتبادل المعلومات . حيث أصبح من السهل نشر المعلومات علي شبكة الانترنت مع زيادة عدد المستخدمين والناشرين ونمو مواقع الويب ، إستخلاص المعلومات من الويب أصبح عملية مرهقة وتستغرق وقتا طويلا نظرا لعدم التجانس والطبيعة غير المنتظمة للبيانات المتاحة علي شبة الانترنت ، التنقيب في الويب يستخدم مختلف تقنيات التنقيب عن البيانات وأدواتها اللازمة لاكتشاف واستخلاص المعرفة المفيدة من وصلات مواقع الويب ، محتوي صفات الويب أو سجلات مستخدمي الويب.

تنقيب استخدام الانترنت هو عملية تطبيق تقنيات استخراج البيانات لاكتشاف أنماط استخدام البيانات على شبكة الانترنت. مثل تخصصات تنقيب البيانات الاخرى، فهو يعرف الاجراءات التي تؤدي الى اكتشاف المعرفة المطلوبة.

هدف هذا البحث هو التنبؤ بتفاعل المستخدمين مع موقع data.gov.uk. المنهجية المستخدمة في هذا البحث عللى النحو التالي: أولا قمنا بتحميل البيانات من موقع حكومة المملكة المتحدة على الانترنت.

ثانيا قمنا باستخدام تقنية العنقدة (التجميع) لتقسيم هذه البيانات الي ثلاثة مجموعات افترضنا تسمية المجموعات (التفاعل العالي ، التفاعل المنخفض ، التفاعل المتوسط) .

استخدمنا بعد ذلك تقنية التصنيف علي هذه البيانات حتى نتمكن من تصنيف بيانات جديدة مثيلة بناءا علي القاعدة التي سيتم استخلاصها من هذه البيانات ونتوقع الي أي مجموعة سيتم اضافة البيانات الجديدة غير المصنفة ، وتوقع التغيرات التي قد تحدث في حجم المجموعات من زيادة أو نقصان .

بعد تطبيق النموذج عن طريق الخوارزمية naïve bayes أوضحت النتائج أن المجموعة الاولي (التفاعل العالي) نتيجتها 283 بينما المجموعة الثانية (التفاعل المنخفض) نتيجتها3 أما المجموعة الثالثة (التفاعل المتوسط) نتيجتها 14.

يمكن الاستفادة من هذه المعلومات كتقديمها كخدمة للزبائن أو تكون معلومات مفيدة للمنافسين في نفس المجال .

1.1 Introduction

Nowadays, the Web has become one of the most widespread platforms for information change and retrieval. As it becomes easier to publish documents, as the number of users, and thus publishers, increases and as the number of documents grows, extract for information is turning into a cumbersome and time-consuming operation. Due to heterogeneity and unstructured nature of the data available on the WWW.[1]

Web mining is the application of data mining techniques to automatically discover and to extract knowledge from web data. [2]

Web Mining is categorized into three categories: Web Structure Mining, Web Content Mining, and Web Usage Mining .

In this research we discusses about web usage mining its techniques, tools, and algorithms.

The concept of clustering and classification is applied in a dataset from United Kingdom government website (data.gov.uk) and analysis it. The results were good.

1.2 Research Problem

The research problem can be viewed from two aspects:

1. How to collect usage mining information.
2. After collecting this information how we get benefits from them. Analysis it to extract useful knowledge, to discover new information or assistance in making a decision.

1.3 Research Objectives

1. Apply clustering technique to grouping data to category (labeled data).
2. Build a prediction model to predict a category of new data and changes that may occur on the categories.(An increase in the size of one of the groups or decrease).

1.4 Research Methodology

In this research we applied clustering and classification technique to the data provided by the British government to its people.

To analysis this data we using rapid miner tool to apply clustering operating, first to grouping this data (data contains four attributes Dataset title, Views, Visits and Resource downloads) according to its similarity into three groups, Rapid miner cleaning the data and replacing missing values before applied algorithm, we used k-means algorithm to grouping data, this algorithm cannot handle polynomial attributes. It converts attributes to numerical. The result of this

algorithm partitioning example set into three groups (clusters), Cluster 0: 939 datasets (pedigree: 0.940), Cluster 1: 12 datasets (pedigree: 0.012), Cluster 2: 48 datasets (pedigree: 0.048), Total number of items: 999. We Assume named this clusters (high-interaction, low interaction and medium interaction group).

To build a predict model to predict the group of new data we used classification technique throw naïve bayes algorithm, its successful classifier based upon the principle of Maximum A Posteriori (MAP). If we insert new data naïve bayes labeled this data according to this rule and predict the changes that may occur on the grouping (An increase in the size of one of the groups or decrease).

The model is applied using naïve bayes algorithm and the result explains: Cluster 0 (height-interaction category) is expected to decrease from 939 to 283 datasets (pedigree expected to increase from 0.940 to 0.943). Cluster 1 (low-interaction category) is expected to decrease from 12 to 3 datasets (pedigree expected to decrease from 0.012 to 0.010). Cluster 2 (medium interaction category) is expected to decrease from 48 to 14 datasets (pedigree expected to decrease from 0.048 to 0.047).

Can take advantage of this information as a service to customers or be useful information to competitors in the same field.

1.5 Thesis Organization

In chapter 2, we talked a comprehensive way about mining in the Web its types, techniques, tools, and named some of the algorithms used for mining.

In chapter 3, we talked in more detail about mining in web usage its process, techniques, tools and some research issues.

In chapter 4, we talked about clustering and classification technique in data mining its defined, types , explain how the algorithm work (only one algorithm for each type) and the comparison between them.

In Chapter 5, we talked about implementation of research tool (as detail).We talked first about the 'rapid miner 'as a tool, then for how to use them in the clustering and classification process, and we discussed the results and Synopsis of the application.

In Chapter 6, Conclusions and recommendation for future work.