Abstract

Nowadays, search tools are crucial to search for information on the Web. However, during this time, the amount of demand information that is available to us on the Web is changing and growing continuously, and those search technologies continually have been pushed to the absolute limit. Therefore, various new techniques have arisen to enhance the search process, and one of them is the analysis of query logs. Query logs have a mechanism to track the history of queries sent to the search engines and the pages favored after a search, among other data.

Classification of web pages is the first step of web page ranking (or we can call it indexing), one of the most ways to achieves indexing process is clustering those pages into groups as per the similarity, whenever the misclassification is too much less, the result will be perfect. Not far away, clustering is a collection of algorithms that divide the data into groups related to each other. For this job, we chose the (Microsoft learn to rank) dataset to achieve the analysis and model building on it, this dataset was designed especially for researches in this field, and it has enormous and different information about the ranking process. Because of a quantity of information, we chose randomly 16015 observations only from MSLR-WEB30K_2 _ fold 1; in this study, according to the ability of our hardware and the algorithms of analysis, some of the algorithms which used in the analysis (determine the optimal number of clusters) can't handle the massive quantity of observations. In this thesis, I will use clustering analysis to improve the web search ranking using comparative analysis between k-means and genetic algorithms as a technique used for this goal.

This process including the clustering analysis to reduce the features using Principal component analysis (PCA) with root main square error as feature reduction technique to compute the error rate and the accuracy of the model result to get the best number of attributes, this process achieved with cross-validation approach using extreme gradient boost algorithm as a training model to estimate the sum of errors during a training operation. After that, we will use various methods to determine the optimal number of clusters to ensure high-quality distribution of the data in the clusters. And to test the result of building models with k-means and genetic algorithms.

الملخص

في الوقت الحاضر، تعتبر أدوات البحث ضرورية للبحث عن المعلومات على الويب. ومع ذلك ومن وقت لآخر، يتغير مقدار المعلومات المتاحة لنا على الويب ويزداد باستمرار، وقد تم دفع تقنيات البحث هذه باستمرار إلى الحد الأقصى المطلق. لذلك، ظهرت العديد من التقنيات الجديدة لتعزيز عملية البحث، وأحدها هو تحليل سجلات الاستعلام. تحتوي سجلات الاستعلام على آلية لتتبع محفوظات الاستعلامات المرسلة إلى محركات البحث والصفحات المفضلة بعد البحث، من بين بيانات أخرى.

يعد تصنيف صفحات الويب الخطوة الأولى في ترتيب صفحات الويب (أو يمكننا تسميتها بالفهرسة)، ومن أكثر الطرق لتحقيق عملية الفهرسة تجميع تلك الصفحات في مجموعات وفقًا للتشابه، فكلما كان التصنيف الخاطئ أقل كلما كانت النتيجة مثالية. ليس بعيدًا، عملية التجميع باختصار، هي مجموعة من الخوارزميات التي تقسم البيانات إلى مجموعات مرتبطة ببعضها البعض. لهذه الوظيفة، اخترنا مجموعة بيانات (Microsoft Learn to Rank) لتحقيق التحليل وبناء النموذج عليها، وقد تم تصميم مجموعة البيانات هذه خصيصًا للباحثين في هذا المجال، وتحتوي على معلومات هائلة ومختلفة حول عملية التصنيف. بسبب كمية المعلومات، اخترنا عشوائياً ١٦٠١٥ ملاحظة فقط من MSLR-WEB30K_2 _ fold 1 في هذه الدراسة، وفقًا لقدرة أجهزتنا وخوارزميات التحليل، حيث أنه لا تستطيع بعض الخوارزميات المستخدمة في التحليل (تحديد العدد الأمثل للمجاميع) التعامل مع الكمية الهائلة من الملاحظات.

في هذه الأطروحة، سأستخدم التحليل التجميعي لتحسين ترتيب بحث الويب باستخدام التحليل المقارن بين الوسائل (k- means) والخوارزميات الجينية (Genetic algorithm) كأسلوب مستخدم لهذا الهدف.

تتضمن هذه العملية تحليل المجموعات لتقليل الميزات باستخدام تحليل المكون الرئيسي (PCA) مع احتساب خطأ المربع الرئيسي للجذر (RMSE) كتقنية لتقليل الميزات لحساب معدل الخطأ ودقة نتيجة النموذج للحصول على أفضل عدد من المجموعات، وقد تم تحقيق هذه العملية باستخدام نهج التحقق المتبادل باستخدام خوارزمية تعزيز التدرج الشديد (eXGBoost)كنموذج تدريب لتقدير مجموع الأخطاء أثناء عملية التدريب.

بعد ذلك، تم استخدام طرقًا مختلفة لتحديد العدد الأمثل للمجموعات لضمان توزيع عالي الجودة للبيانات في المجموعات. واختبار نتيجة بناء النماذج وفقا لمعايير تقييم نماذج التجميع المتعارف عليها.