香港浸會大學
HONG KONG BAPTIST UNIVERSITY

HKBU Scholars

MASTER'S THESIS

Analysis of internet image search performance

Wang, Xiaoling

Date of Award:
2011

Link to publication

General rights
Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

Download date: 05 Dec, 2021

Analysis of Internet Image Search Performance

WANG Xiaoling

A thesis submitted in partial fulfilment of the requirements

for the degree of

Master of Philosophy

Principal Supervisor: Prof. LEUNG Clement Ho Cheung

Hong Kong Baptist University

July 2011

Abstract

With the rapid advancement of digital cameras and the Internet, a large collection of digital images can be easily created, shared and distributed, using not only computers, but also numerous other portable digital devices. As digital images have become fully ubiquitous in our lives, searching for the relevant image objects has thus become an important activity. It is desirable to be able to find a solution to searching for images effectively and efficiently. However, many raw images are constantly uploaded without meaningful text labeling or with few words based on the interests of the owner, which is not always reliable and informative. The Internet image search engines based on keywords retrieval, such as Google, Yahoo and MSN, tend to return a large number of images which the engines consider to be relevant, and such pool of results generally is very large and may be regarded as effectively inexhaustible. While the images are presented as relevant, it is normally true that many of them are actually irrelevant, and that the distribution of relevant images over the returned results is non-uniform. Therefore, to predict the distribution of relevant images for the Internet image search engines has become a critical and urgent issue.

The prediction of the relevance for individual images is generally difficult since it only takes on binary values and therefore tends to oscillate randomly between relevance and irrelevance with insignificantly noticeable trends. Increasing the range of possible values is necessary to enhance the prediction ability and it is advantageous to accumulate the aggregate relevance for larger groups of images in a sequential manner. Here, we present a partition approach to the number of relevant images. This will involve

appropriately grouping the random binary sequence into non-overlapping groups and converting it into a form which makes them more amenable for prediction.

Here, we present a Regression model and Markov Chain model for predicting Image Search Engines (ISEs) behaviour. The framework of our approach is initially to design a set of benchmark queries, and then the distribution formula or Markov Chain model will be able to fit the experimental observations by using appropriate parameters and so providing a mathematical description of an empirical process. These two models are particularly effective for the prediction of relevant images for image search engines. The experimental results show that they are able to give good and robust predictions of search engine performance. In addition, the results of this research can have a direct bearing on search engine design to provide informative guidance to users on the retrieval of relevant images, and allows the users to optimize their strategy in the recovery and discovery of images.

After developing the Regression model and Markov Chain model, the estimation of recall becomes easier. In addition, an approach called tagged relevant images, which mainly adopts hypergeometric distribution theory, is developed to estimate recall. The experimental results show that the result is slightly inferior to that of the corresponding estimation methods to create a regression model and Markov Chain model.

Keywords: image retrieval, linear regression model, moving average, exponential smoothing, Markov Chain model, recall, tagged relevant images method

Table of Contents