SCENES DETECTION AND RECOMMENDATION
ACCORDING TO WEB PHOTOGRAPHS WITH GPS
INFORMATION

Hao-Chen Wang
Chang Gung University
No.259, Wunhua 1st Rd., Gueishan Township, Taoyuan County 333, Taiwan (R.O.C.)

Hsien-Tsung Chang
Chang Gung University
No.259, Wunhua 1st Rd., Gueishan Township, Taoyuan County 333, Taiwan (R.O.C.)

ABSTRACT
In this age of fast-growing markets for tourism, ever shortening IT development life cycles and ubiquity of digital cameras, the sharing of travel photos on web albums, online communities and personal blogs has become increasingly popular (e.g. through Flickr or Wretch web albums). To mine relevant travel information from the vast troves of photographs on the Internet, the most critical challenge is to analyze and filter the “noises” of travel photographs. We accomplished and implemented this over four stages, and restricted the range of photographs to those taken in Taiwan. Other than displaying highly popular scenic spots, we also tabulated the current popularity of the scenic spots with rankings created with Tourist PageRank. We further attempted to verify the validity of our system with a number of experimental approaches.

KEYWORDS
Scenes Detection, Scenes Recommendation, EXIF, Page Rank

1. INTRODUCTION
1.1 Research Background and Motivation

According to the World Travel & Tourism Council (WTTC) [1], the global tourism industry currently accounts for 9.4% of the global gross domestic product (GDP), and is projected to increase to 9.5%, or about 10.5 trillion U.S. dollars, in 2019. In addition, statistical data in 2009 showed that international tourists visiting the Asia-Pacific region in 2007 grew 10.5% in comparison to 2006, a significant increase. The number of tourists in 2008 was approximately 180,000,000, up just 1.6% from the previous year, but the global tourism industry is nevertheless an important sector of growth for the global economy.

In the past, digital cameras were capable only of generating an electronic image file. Nowadays, most digital cameras are further equipped with the Exchangeable Image File Format (EXIF) photograph information capability. EXIF is an exchangeable image format, specifically designed for incorporating digital camera photograph information. The Japan Electronic Industry Development Association established the standards for version 1.0 in 1996, and the latest official standards currently stand at 2.2. It can also be tagged onto JPEG, TIFF, RIFF and other document formats. Figure 2 shows the basic EXIF structure of the uncompressed files. The GPS IFD values can be read on digital cameras supporting GPS, and when one travels with GPS satellite navigation, one can embed and record GPS information in a photograph by artificial means [3]. In Figure 3, with the use of the relevant applications, one can see a list of information that can be recorded by photograph EXIF, such as GPS information.
There are currently several online albums communities that enable display of the EXIF information. For example, Picasa provides information ranging from the model of camera used, ISO, latitude, longitude, etc. From version 3 of the wildly popular software ACDSEE, the Metadata field information will also be displayed if the photograph contains EXIF information.

### 1.2 Research content and contribution

In accordance with our research motivation, we started from travel photograph collections of the general public and the tourism information embedded in the photographs. From the photographs and embedded EXIF information obtained from the web, we located scenic spots by means of filtering noise detection. Since the popularity of scenic spots will vary with time, we isolated scenic spots of interest by automatic detection, and identified trends in the rankings thus obtained.

Our starting point was the EXIF information embedded in the photographs. When crawling for photographs on the web, we also captured other relevant information and employed analysis filtration accordingly, as outlined in the following chapters. We conducted exploration of scenic spots and provided tourism-related recommendations, using the Tourist PageRank calculation method to compile relevant rankings.

PageRank [5], also known as page rank, was an algorithm invented by Sergey Brin and Larry Page, the founders of Google. Its main application is search engine technology, for indicating the relevance and importance of web pages. Based on the hyperlinks between web pages, one can calculate a page’s rank using the core concept of the vote for the hyperlink. The complete equation is:

$$\text{PageRank}(p_i) = \frac{q}{N} + (1-q) \sum_{p_j} \frac{\text{PageRank}(p_j)}{L(p_j)}$$

where p1, p2, ..., pN are the studied pages, M (pi) is the number of page links into pi, L (pj) is the number of page links out of pj, and N is the number of all pages. q is known as the damping factor, and it represents the probability that a random visitor will continue on to other pages of a website after he or she has browsed a particular page.

The PageRank of individual pages is determined by the voting scores, and the number of external links to an individual page is representative of its scores. For example, if someone links to your page, it represents his or her vote for you. However, the vote scores will also depend on the global total of external links. If the PageRank value is 5, and 10 persons have voted (10 external links), then score is only 0.5 points. Therefore, the greater the number of linked pages, the higher the value of the PageRank.

### 1.3 Paper structure

This paper is divided into five chapters. Chapter I consists of the foreword; Chapter II details our approach; Chapter III describes the experimental design, analysis and results; Chapter IV presents our survey of the relevant literature; Chapter V outlines the advantages, limitations and possible applications of our study.
2. SYSTEM DESIGN AND IMPLEMENTATION

2.1 System Architecture

Figure 4 shows our system structure. First, we obtained photographs published by users on the Internet, with the sources encompassing the entire network (such as Internet photograph albums, groups and blogs, etc.) as well as more concentrated sources such as Flickr [6] and other online communities for sharing photograph albums. We next wrote a program to gather data according to techniques fine-tuned over the years in our laboratory. We analyzed the EXIF information of each photograph and filtered “noises” in the photograph information using a set rules we formulated to determine the user’s travel behavior sequence. Next, we used the processed information to identify the scenic spots frequented by the users. Finally, we ranked and presented the results, and provided relevant recommendation.

2.2 Information Processing

We used the Data Gathering function to capture photographs on the Internet to our servers, and recorded the source URLs of the photographs on the Internet, to enable follow-ups for users. This way, a user interested in certain scenic spots will not need to go through the search engine a second time to unearth more information. The data filtering and analysis consisted of four stages. First, we obtained the relevant photograph related information. Second, we determined the chronological order of the travel. Third, we conducted sequence analysis of the scenic spots. Fourth, we ranked the scenic spots.

2.2.1 To obtain information about photograph

We captured the location of a photograph on the server, using the ReadExif function to obtain the EXIF information, and recorded the EXIF fields: "photography time, file name, camera model, latitude, longitude, version>, e.g. <Jun 28, 2008, P6272809.jpg, OLYMPUS IMAGING CORP., latitude 50.444469 °, longitude 18.855607 °, 2.2.0.0>, using "EXIF_ field name" to represent each field. In the program flow diagram shown in Figure 5, we first determined the EXIF information of the photograph. If the information did not exist, we then set the version of field to 0, to resolve the different versions of the format. If the EXIF_ photography time was not present, we would then select the earlier of the community site file upload time recorded or the time the photography was said to be taken from the website.
2.2.2 Determine personal travel sequence

During this stage, we analyzed the results generated from the previous stage. We defined the size of the tourism sequence to be 1 ~ n, with time in days as the unit of $\alpha$ and distance in km as the unit of $\beta$.

We performed the following three steps for all the information in the table. The first was cluster, and we filtered the noise with three conditions. The first condition was to ensure that there was one and only one unique user ID for each user, with the assumption that there existed no external factors interfering with the experiment (e.g., the use of more than one camera for a particular trip). With the user ID in place, we obtained the user's name automatically with the program we wrote. Its capture was based on the following three criteria: source URLs from the same web directory, description on the originating webpage (photograph album of the user ID), hyperlinks on the originating webpage.

The second condition: The same group of photograph taken over more than $\alpha$ filters days (set $\alpha = 3$). Condition 3: Filter the distance of the photograph by more than $\beta$ km ($\beta = 200$), with the distance commonly assumed for travel in the western half of Taiwan as an example. The distance from Taipei to Pingtung is about 400 km, a distance that can be used to determine the scope of tourism in northern Taiwan or southern Taiwan. The second step was the cluster of the previous step, according to the chronological order, with each cluster of information used to save a sequence. The third step was to analyze the travel sequence, a sequence that did not exceed more than two hours for each photograph, failing which we undid the process and reverted back to step one.

This way, we can infer the individual behavior travel sequence, a sequence comprising of several data points, each containing a photograph and the geographic location information, a sequence that was effectively a record of the user's travel through time. Although the use of the time at which the photographs were taken to reenact the user's travel itinerary leaves some margins of error with respect to time, we do not believe they have significant impact on our experimental results. Should we be unable to read the page by EXIF and the photograph's photography time, we would consider photographs sharing the same directory to be taken at the same time. For the above mentioned filter and the method for determining travel, we do not foresee significant effects on our experimental results.

2.2.3 Sequence analysis of tourist attractions
In this section, we based our analysis on the travel sequence determined in the previous section and determined the distance from the GPS information of each photograph. The distance of $\gamma$ km ($\gamma = 1$) within the photographs was chosen based on the assumption that few visitors will traverse a distance exceeding one km from the scenic spot. We then analyzed the name of scenic spots and obtained the names of the scenic spots from the web as well as the user’s description.

We retained five candidate names for each scenic spot (e.g., Truku, Taroko, Taroko National Park, Hualien County, Taiwan) and compiled statistical rankings for each candidate name of a scenic spot. When all the scenic spots statistical ranking was complete, we then filtered the name of the scenic spots.

According to the statistical rankings of the scenic spots’ names, we selected the three most frequently appearing ones from the five candidates as the name of the scenic spots (e.g.: Truku, Taroko, Taroko National Park). This way, we obtained a more popular but not overly broad range of names for the scenic spots.

For names in other languages, we selected visitors speaking the respective languages, obtained descriptions of the photographs and titles on the Internet, and then chose the three spot names which appeared the most frequently. We then filtered in the same manner as described above.

![Figure 8. The relationship between tourist attractions and individual behavior](image)

### 2.2.4 Ranking tourist attractions

We next applied the commonly used method for computing search engine rankings, as described in the introduction, to determine our ranking of the scenic spots — Tourist PageRank. First of all, we assumed that the users took pictures only at locations considered scenic spots, and that some of these photographs also depict the users in their own homes or personal settings or some non-tourist spots. We filtered the noise by means of a ranking score threshold.

As formula I and formula II in the following, $\text{SR}(X)$ represents the scores of the scenic spots, and $\text{TR}(Y)$ represents the scores of personal travel behavior, with the relationship as shown in Figure 8. We treat a scenic spot as a web page and the personal travel behaviors as links those are similar to Page Rank algorithm for ranking the web pages.

**Formula I:**

$$ \text{TR}(Y) = \sum \text{SR}(X) $$

**Formula II:**

$$ \text{SR}(X) = \sum \text{TR}(Y) $$

In order to match real situation of scenic spots, we introduce the number of photos as $P(X)$ to indicate the total number of photos that tourist took in the scenic spot. We think that if a scenic spot is amazing for a tourist and he would take more photos than one scenic spot which is bored. So the value $P(X)/\text{Numberof}(Y)$ indicates the attraction score for a scenic spot according to the average number of photos when a tourist visit that spot. It is similar to the content score for a web page. We modified the formula II to formula III as following.

**Formula III:**

$$ \text{SR}(X) = \left( \sum \frac{\text{TR}(Y)}{\text{Numberof}(Y)} \right) + \frac{P(X)}{\text{Numberof}(Y)} $$
We applied the above formulas and calculated the spots scores $SR(X)$ and personal travel scores $TR(Y)$ iteratively. With the four stages of analysis and filtering as defined in Chapter 2.2, we finally generated a ranking of the spots. We filtered the scenic spot scores which were lower than 100 and presented our list of recommended scenic spots based on the scores in the ranking.

3. EXPERIMENTAL DESIGN AND ANALYSIS

Our study made use of four servers to gather web photographs, and then a personal computer for performing the analysis to obtain and filter for the required information, with the final results stored on a server.

In our experimental set-up, the range of tourist attractions was restricted to Taiwan. However, we did not gather photographs only from websites based only in Taiwan, for the simple reason that many foreign tourists took photographs in Taiwan on their travels, in order to enhance our multi-language support. To enable users to obtain more information on the attractions, we can increase the volume of information covered by the query results. This will also facilitate future experiments to further expand the range of exploration sites worldwide.

In the experimental methods section, we grouped the results of our analysis into two categories. The first category caters to the assessment of attractions for which we offered relevant information. We attempted to verify from the users the accuracy of the information provided. The second category was mining attraction assessment. We attempted to better understand this kind of analysis and evaluation, how we could validate the accuracy of our system by comparison with real attractions. We will describe in details the experimental method as follows.

3.1 Attractions Information Assessment

For this first category, we validated the veracity of other resources provided by the webpage, including tourism-related information described in previous chapters. For collecting data, we created web pages in Q & A mode, with several buttons for users to evaluate whether the information that we compiled was accurate.

3.2 Mining Attraction Assessment

For this assessment component, we applied the Blind Test to compare the validity of our recommendations with those from travel sites providing similar recommendations. Here, we randomly selected pages from these sites in random display modes for side by side comparisons with our own, and deliberately concealed distinguishing features of the websites to prevent users from forming preconceived impressions, for a fair appraisal.

Other than displaying the popular attractions, we also attempted to determine the popularity of the attractions. We denoted the more popular attractions with large icons, and determine their popularity with the number of clicks, for comparison and reconciliation with the results obtained from our data.
In another assessment of our system, we employed traditional electronic questionnaires to determine the level of satisfaction with the attractions in the localities, with questions on the attractions that interest the local populace and whether our system had successfully identified them.

Finally, we collected various bodies of information on well-known attractions such as Yang-ming-shan, Alishan, etc. to compare with our system and test the credibility of our system.

4. RELATED WORK

Y. Yang [8] applied the OWL approach in the collection and establishment of tourist information, where OWL is the acronym for Web Ontology Language. OWL is intended for use in cases where the data is process by computer applications rather than by human. In other words, it must be understood by the machine language. OWL can be used to clarify the meaning of terminologies, as well as the relationship between terminologies. When it comes to expressing meanings and semantics, OWL enjoys more means of expression compared to XML, RDF, and RDF-S type syntaxes. Therefore, to express a language readily understandable to computers in the web which is stronger than the above, which was designed primarily to process the contents of information, we took advantage of a feature for network and employed this method on a large number of travel sites to parse and integrate information. They used an OWL editor for graphical representations of concepts, where a line represents the relationship between concepts, allowing users to easily establish the relationship between self-concept and semantics. In the future, we hope to apply these methods to achieve automated deployment of travel information, but the syntax of such sites is still rather small, meaning the information thus obtained is very limited.

In 2009, A. Popescu [9] et. al attempted to match photographs supplied by users and the names of attractions in the photographs. They analyzed photographs whose file names were changed by the user, or web pages where the photo was described by the title and context. They made use of Wikipedia to obtain the names of attractions and construct a database, and then leveraged the vast resources of Wikipedia to obtain their counterparts in over 10 foreign languages, thereby creating geographic content with vocabularies from multiple language to increase their matching range.
In Figure 12, the attractions on the left had 100 or less travel photographs whereas those on the right had 1000 or more photographs. It can be seen that there existed more famous attractions from before, and they attracted more visitors. We can identify several less popular attractions from this sample collection. In contrary to the methods of this paper, Wikipedia may not have included all the attractions, and those included are likely to be the most widely known but not necessarily the most popular attractions. In addition, the validity of a sample collected by manually altering nicknames and information on webpages is debatable.

Y. Zheng [10] also attempted to mine interesting sights by recording user travel behavior with GPS tracking. First, they defined the destinations, which contained two parameters: time thresholds and distance threshold, to determine the tourist attractions. They subsequently re-used the density of cluster algorithms to eliminate the noises in the information, with the information collected clustered at different levels. Here, the noise referred to areas where the cluster density was relatively sparse areas, and where TBHG was used to modulate the travel sequence.

After the completion of the above, one would obtain the clusters at different levels, as shown in Figure 13 on the right. One can then apply the HITS inference module to directly link the users and attractions. Here, the authors used two parameters. Authority represents the scores for the attractions, which is the total of the scores for hubs pointing toward this attraction. The higher the scores, the larger the number of high-frequency users. Hub represents the scores of the user, which is the sum of the scores of attractions it points toward. A higher score indicates that the user has visited many high-frequency tourist attractions. One then iteratively computes the scores to rank the attractions and users. This paper made extensive use of GPS log, and since not everyone owns such a device, it would be more difficult to gather travel information effectively. On the other hand, the use of the HITS method for score is more easily manipulated, since a user may travel frequently and thus has a higher score, but he or she may not necessarily visit the more interesting attractions.

5. CONCLUSION

We have effectively exploited photographs that recorded the sources of information and mined relevant travel information, using the analysis and filtering method as outlined above to remove the noise in the tourist attractions information. Lastly, we used the Tourist-PageRank to remove homes, offices and other spaces for personal activities, and presented a system that ranks scenic spots in order of popularity as well as generates recommendations for users.

The source of our data is the entire body of photographs on the Internet around the world, but we limited our range of tourist attractions to Taiwan by restricting exploration out of range, in order to facilitate a proper assessment of the feasibility of our system. We gathered data at the same time, to provide information in different languages in terms of attraction names, descriptions from foreigner, etc. and we will obtain more alias names for popular attractions to increase the scope of the user query.
user with an interest in a particular tourist attraction can access photographs from a direct link on our system to the web page where the photographs originated without the use of additional search engines, websites, tools to obtain more information on the tourist attractions. Finally, we designed experiments using previously formulated methodologies to effectively get the system to automatically obtain the name of tourist attractions and verify feasibility and accuracy of information.

In the future, we hope to further expand the scope of our study to cover tourist attractions around the world, enabling more and more users to plan their budget travel accordingly. We intend also to study how best to help recommend planning of travel itinerary to cater to the needs of different individuals, a research topic we believe to be of utmost importance to the travel industry.

ACKNOWLEDGEMENT

The author would like to express his gratitude to the numerous faculty members who have provided him with invaluable advice and guidance, as well as the many co-workers in the laboratory and friends who have offered him words of encouragement.

REFERENCES

透過具有GPS資訊網路照片之景點偵測與推薦
Scenes Detection and recommendation according to web photographs with GPS information

指導老師：張賢宗 博士
研究生：王皓塵

中華民國 九十八 年 十二 月
目 錄
第一章 序論........................................................................................................1
  1.1 研究背景與動機....................................................................................1
  1.2 研究內容與貢獻....................................................................................4
  1.3 論文架構................................................................................................6
第二章 系統設計與實作....................................................................................7
  2.1 系統架構................................................................................................7
  2.2 資訊處理................................................................................................8
第三章 實驗設計與分析...................................................................................17
  3.1 景點資訊評估......................................................................................17
  3.2 景點資訊評估......................................................................................18
第四章 結論與未來展望...............................................................................21
第五章 相關研究............................................................................................22
參考文獻........................................................................................................24
摘要

在這世界觀光旅遊、資訊發展及數位相機普及的時代，將旅遊照片分享於網路相簿、社群和部落格，已越來越為盛行（如網路相簿 Flickr 使用者數量或國內知名的無名小站會員數等等），在網絡如此廣大且豐富的資源下，我們取得網路上照片所帶來的旅遊資訊開始著手，而如何分析及過濾這些旅遊照片中所帶來的雜訊為其中最主要的挑戰。我們使用四個階段來分析及過濾這些照片資訊，並將系統實做出來，實驗範圍為我們所在的台灣，表示所探勘出的景點區域為台灣景點，系統除了呈現探勘出的旅遊景點外，更進一步的將景點以 Tourist PageRank 的方式，將景點與使用者不斷以迭代法方式，計算旅遊景點目前熱門程度的排名。藉由我們所設計的實驗方法，能有效的證明我們的系統可信度與探勘出的旅遊景點準確率以及景點熱門資訊的可信度。
第一章 序論

1.1 研究背景與動機

世界觀光旅遊委員會（WTTC）指出目前全球觀光產業規模預計占全球國內生產總值（GDP）的 9.4%，更預計在 2019 年將會提升至 9.5% 約為 10.5 兆美元，此外在 2009 年的統計資料顯示了我們所在的亞太地區，國際觀光客在 2007 年造訪亞太地區較 2006 年成長了 10.5%，有相當大幅度的成長，2008 年約有 1 億 8,800 萬人次的觀光客，僅管僅僅成長 1.6%，但對於觀光旅遊產業部分還是在全球經濟成長中扮演重要的角色。

此外，隨著科技多年的發展，科技產品已經在日常生活中各個層面提供了無數的便利，由圖 1.1 我們可看出數位相機銷售價格從目前至未來幾年已經不斷的下跌，由於數位相機的普及，人們便越來越容易藉由數位相機來記錄個人的日常生活與旅遊活動。

![圖 1.1 數位相機平均銷售價格趨勢預估](image)

過去數位相機單單僅提供一張電子圖檔，現在大多數的數位相機...
都有附加 Exchangeable Image File Format (EXIF) 没有照片資訊，EXIF 是一个可交换图像格式，专门設計用来表示數位相機的照片資訊，從 1996 年由日本電子工業發展協會制定的 1.0 版本，發展至今已有 2.2 官方標準，更可以附加於 JPEG、TIFF、RIFF 等文件之中。圖 1.2 為 EXIF 未壓縮的檔案基本結構，而其中的 Value of GPS IFD 結構可以利用支援 GPS 的數位相機或出外時搭配衛星導航來嵌入 GPS 資訊於照片，更可以利用人工的方式將 GPS 資訊記錄於照片內，圖 1.3 為使用相關應用程式來瀏覽照片 EXIF 資訊的範例，可以很明顯的看到它所記錄的 GPS 相關資訊。
目前已有許多網路相簿社群提供 EXIF 相關資訊的解讀，以 Picasa 為例，它們所提供的照片資訊包括模型、ISO、緯度、經度等等，另外像著名看圖軟體 ACDSEE 在 3.0 開始，如果照片含有 EXIF 資訊，也會另外提供 Metadata 欄位的資訊。

1.2 研究內容與貢獻

藉由上述研究背景與動機，我們從人們的旅遊照片收集開始著手，加上旅遊照片所帶來的旅遊資訊，在這透過照片所在網頁及內嵌 EXIF 所帶來的旅遊照片資訊，藉由過濾雜訊的方式以達到景點偵測。由於景點的產生、消失、好壞隨著時間的流逝一直在改變，我們透過自動偵測的方式，來挖掘出更有趣的旅遊景點，也可藉此挖掘的結果得知目前景點趨勢的排名。

我們以研究背景中所介紹的照片資訊 EXIF 為基礎，並在系統抓取照片的同時，將照片所在網頁的其他資源一併撿取至我們的資料。
庫，透過上述來源所獲得的資訊，使用我們後續章節所定義的分析過濾方式，探勘出旅遊景點，並進一步提供旅遊相關資訊，再利用 Tourist PageRank 計算的方式，對探勘出的景點加以排名，將本系統實作出來。

Pagerank 又稱為網頁排名，它是 Google 創始人所發明的技術，主要應用在搜索引擎技術上，用來表示網頁相關性與重要性，根據網頁之間相互的超連結來計算網頁排名技術，核心的概念為讓連結來投票，假設有 n 個頁面連結至你的頁面，那麼所獲得的分數（PageRank）等於個別連結頁面的 PageRank 除以個別的對外連結數。

完整的等式：

\[
PageRank(p_i) = \frac{q}{N} + (1-q) \sum_{p_j} \frac{PageRank(p_j)}{L(p_j)}
\]

\(p_1, p_2, ..., p_N\) 是被研究的頁面，\(M(p_i)\) 是鏈入 \(p_i\) 頁面的數量，\(L(p_j)\) 是 \(p_j\) 鏈出頁面的數量，而 \(N\) 是所有頁面的數量，\(q = 0.15\)（\(q\) 被稱為阻尼係數（damping factor），其意義是，在任意時刻，假想的隨機瀏覽者停止在某頁面後繼續瀏覽的機率。

個別連結頁面的 PageRank 就是它們手上所擁有的投票分數，個別的對外連結數就代表了它們自己的分數均分成幾份投票出去。（例如：別人連結至你的頁面，代表它投了票給你，但是這票的分數，就取決於它把票給幾個人，如果它手中的 PageRank 值為 5，而它把票...
投給 10 個人 (十個對外連結)，那麼你的頁面所獲得的分數就只有 0.5）。所以，連結的頁面越多，PageRank 就越高。

1.3 論文架構

本論文共分為五個章節，第一章為序論，介紹研究背景與動機、內容與貢獻及論文架構；第二章為系統設計與實作，詳細的介紹我們所建置的系統架構及資料探勘、過濾雜訊、排名計算等方法，第三章實驗設計與分析，介紹各個實驗所帶來的意義與設計方式，並以實驗結果做到系統評估；第四章述說論文結論與未來可探討的研究方向。最後在第五章呈現我們所瀏覽過的相關文獻。
第二章 系統設計與實作

2.1 系統架構

圖 2.1 系統架構

圖 2.1 顯示了我們的系統架構，首先在網路上取得使用者所發佈的照片，照片來源為網路上所有照片（如網路相簿、社群和部落格等等），較為集中的來源如 Flickr 及其他網路社群相簿。藉由本實驗室資料抓取技術為基礎，以本系統為考量撰寫程式進行資料抓取，接著進行照片 EXIF 資訊解析，以及照片所在網頁相關資訊的收集與 EXIF 資訊進行比較，以決定對於該照片的資訊解析，將解析完的資訊使用我們所定義的規則來過濾照片所帶來的雜訊資訊，並判斷出使用者旅遊行為序列，再使用處理過的資訊進行解析，探勘出使用者的旅遊景點，最後將結果使用網頁的方式呈現給使用者，我們呈現的資訊除了經由每個階段所探勘出的景點外，並做到排名及提供相關資訊。
2.2 資訊處理

我們使用所撰寫的 DataGather 函式，將網路上的照片抓取至我們的伺服器，並將照片所在網頁記錄起來以利後續步驟使用，以及提供照片來源網址給使用者，如此一來使用者對於有興趣的旅遊景點便不用重新使用搜尋引擎尋找欲查詢的資訊。接著便進行資料過濾與分析，在這裡分為四個階段，第一階段取得照片相關資訊；第二階段判斷個人旅遊序列；第三階段分析旅遊序列景點；第四階段旅遊景點排名。

2.2.1 取得照片相關資訊

![流程圖](image-url)

圖 2.2 取得照片相關資訊流程
我們將伺服器裡所抓取的照片，使用 ReadExif 函式取得照片 EXIF 資訊，在此取得並記錄 EXIF 欄位：〈拍攝時間、檔案名稱、相機型號、緯度、經度、版本〉，ex：〈Jun 28, 2008、P6272809.jpg、OLYMPUS IMAGING CORP.、北緯 50.444469 度、東經 18.855607 度、2.2.0.0〉，以 “EXIF_欄位名稱” 來表示每個欄位。在程式流程圖 2.2，先判斷照片 EXIF 資訊，若不存在則版本欄位設為 0，以解析不同版本格式；若無 EXIF_拍攝時間，則以社群網站檔案上傳時間記錄與 Web 本文所描述的拍攝時間，選取較早時間為拍攝時間；

![流程圖 2.2 取得照片 EXIF 資訊](image)

**圖 2.3 取得照片經度緯度資訊**
在經度及緯度欄位則利用 `gpsGet` 函式來判定照片 GPS 地理位置，在 `tempName` 陣列裡，我們儲存三個較為頻繁出現的地理名詞，向搜尋引擎發出查詢，以陣列[0]為例：“tempName[0] GPS”，使用陣列0所儲存的地理名詞 + 空格 + GPS 發出查詢。

藉由上述流程及函式，儲存每張照片資訊至照片資料表，資料表欄位為：<ID、照片名稱、經度、緯度、拍攝時間、相機型號、EXIF版本、所在網頁位址>，ID 為新增一筆記錄時，系統自動產生序號。

2.2.2 判斷個人旅遊序列

此階段我們以上一階段所產生的結果進行分析，在此定義旅遊序列大小為：1 ~ n 個，時間 α 以天為單位、距離 β 以公里為單位。

![判斷個人旅遊資訊流程](image)
將照片資料表裡的所有資料，進行三個步驟，第一步驟：分群，在此藉由三個條件來過濾雜訊，條件一：判斷使用者ID為同一使用者，在此假設沒有外在因素干擾實驗（如：使用許多相機記錄當次旅遊行為）。使用者ID根據照片所在網頁，藉由我們所撰寫的程式自動取出使用者的名稱，其擷取方式如下：根據來源網址同一目錄名稱（網路相簿http://www.wretch.cc/album/使用者ID）、本文描述（使用者ID的相簿）、網頁內容含有超連結（「用照片記錄XXX的眼中世界」）三方面來擷取。條件二：將同一群的照片過濾拍攝時間超過α天（設α=3）。條件三：過濾照片距離超過β公里（β=200），在此的距離假設以台灣西半部為例，從台北到屏東距離大約為400公里，判斷旅遊範圍為北台灣或南台灣。第三步驟解析旅遊序列，一序列中照片相互間隔不超過兩小時，否則將序列拆開回至步驟一。

透過上述方式，得到個人行為旅遊序列，一序列由多筆資料所組成，每筆資料為一張照片與所表示的地理位置資訊，一序列為同一使用者當次的旅遊記錄。儘管透過使用者的拍攝時間來解析個人旅遊序列，但對於時間資訊的準確性，我們認為對於實驗結果沒有影響，假設我們無法藉由EXIF及照片所在頁面解讀拍攝時間，此時，我們將同一時間及目錄下所得到的照片，解讀為同一時間拍攝的照片，對於前述的過濾及判斷旅遊的方式，不會造成實驗的影響。
2.2.3 分析旅遊序列景點

依據前一節所設置的旅遊序列，對序列中每張照片的 GPS 資訊來判斷其距離，將距離 \( \gamma \) 公裡 \((\gamma = 1)\) 以內的照片，判斷為一旅遊景點。我們認為旅客在景點地標活動範圍不會超過一公里，接著分析旅遊景點名稱，藉由該景點所在網頁本文獲得景點名稱，以及從使用者描述中取得景點名稱。

每一個旅遊景點，保留五個候選景點名稱（如：Truku、太魯閣、太魯閣國家公園、花蓮縣、台灣），接著對於每一個旅遊景點的候選名稱進行統計排名，當所有的旅遊景點統計排名完成時，便進行過濾景點名稱的動作。

我們根據景點名稱的統計排名，選擇五個候選景點名稱裡，三個較為不頻繁出現的景點名稱，做為該景點名稱（如：Truku、太魯閣、太魯閣國家公園），如此一來我們便能取得較為通俗但又不是大範圍的景點名稱。

多國語言名稱部分，我們抓取使用其他語言遊客對於該照片所在網頁的本文描述及標題來取得三個較為頻繁出現的景點名稱，其過濾方式同上述方法。

2.2.4 旅遊景點排名

在景點排名部分，我們以序論所介紹搜索引擎常用的排名計算方
式為基礎，來訂定我們的旅遊景點排名—Tourist PageRank，首先我們假設使用者所拍照的地點皆為旅遊景點，當然這些旅遊景點裡，包含了使用者在自家及個人場所的自拍照片或一些不在景點而拍照的旅遊照片，我們在此藉由此排名分數的門檻來過濾篩選此類雜訊。

對於每一個候選旅遊景點，對於一個 User ID 限制一個旅遊景點照片 3 張以上，接著對於景點與個人進行評分，我們在裡面使用幾個參數，TourSoc 表示旅遊景點分數、Hubs 表示個人旅遊行為分數，其相互關係如圖 2.4 所示。

![圖 2.4 旅遊景點與個人行為相互關係](image)

圖 2.4 旅遊景點與個人行為相互關係
取出旅遊景點

公式一: 個人當次旅遊行為對於該景點分數

公式二: 由於拜訪景點所增加的個人旅遊行為分數

公式三: \( \Sigma \) 到過該景點的個人分數

公式四: 景點對於到訪過的使用者所給予的分數

圖2.5 旅遊景點分數計算

公式一: 個人當次旅遊對於該景點的分數 HubTour

\[
\text{HubTour} = \frac{\text{本次旅遊照片數}}{\text{過去拜訪次數} \times (1 + \text{每年拜訪頻率})}
\]

每年拜訪頻率我們以一年為單位，例如某位使用者拜訪某景點兩年中到訪三次，則每年拜訪頻率 \( = \frac{3}{2} = 1.5 \)。

公式二: 個人拜訪旅遊景點所增加的分數

\[
\text{Hubs} = \frac{(\text{Hubs} + \text{HubTour})}{(\text{拜訪過的景點個數} + 1)}
\]
公式三：旅遊景點分數

\[ \text{TourSoc} = \sum \text{到過該景點的人 Hubs} \]

公式四：景點對於到訪過此景點的人給予的回饋分數

\[ \text{Hubs} = \text{Hubs} + \text{該景點的照片數} \times \text{使用者拜訪該景點的頻率} \]

在公式四，該景的照片數為所有到過此景點的旅行者所拍攝的照片數總和；使用者拜訪該景點的頻率，我們以一年月份為單位，例如某位使用者兩年拜訪該景點兩次，則拜訪該景點的頻率 \( = \frac{2}{24} = 0.08 \)。

經由上述四個公式，不斷的使用迭代方式，反覆的計算其景點分數(TourSoc)與個人旅遊分數(Hubs)。

我們經由 2.2 開始所定義的四個分析過濾階段，最後產生了景點的排名，我們在此過濾景點分數 < 100 的旅遊景點，最後在推薦景點呈現部份將採用此景點分數排序來呈現。
第三章 實驗設計與分析

我們一開始使用四台伺服器進行網路照片抓取，接著使用一台電腦進行第一階段的圖片相關資訊分析，取得所需的相關資訊，使用一台電腦對解析出來的相關資訊進行過濾，並將結果放置伺服器，以 Web 方式呈現給使用者。

目前所訂定的實驗範圍為探勘出的旅遊景點所在範圍結果為我們所在的台灣，在這裡所收集的照片不見得僅位於台灣的網站，更必須收集許多國外旅遊者在台灣所拍攝的照片都涵蓋在內，以增進我們對於多國語言支援，讓使用者欲查詢景點時，增加其查詢結果涵蓋的資料數量，此外，便於未來將實驗更進一步的擴大至全世界的景點探勘。

實驗方法部分我們根據分析結果分為兩類，第一類為景點資訊評估，對我們提供的景點相關資訊，想要從使用者使用上得知所提供資訊是否準確；第二類為景點探勘評估，希望藉由此類的分析與評估得知利用本系統所探勘出的景點，與真實景點相比得到系統結果的準確率。我們將在下方詳細的介紹實驗方法。

3.1 景點資訊評估

在一類景點資訊評估方面，我們對於網頁所帶來的其他資源評估其準確率，其項目包含了在前面幾個章節欲呈現的旅遊相關資訊，
收集的方法採用網頁常用的問答方式，在呈現景點資訊的同時，在資訊下方採用讓使用者評價的方式進行評估，評估所收集的資訊是否準確。

3.2 景點探勘評估

第二類別景點探勘評估部分，首先使用 Blind Test 的方式，根據我們所收集的幾個景點推薦網站，隨機選取所收集的頁面與我們的網站推薦頁面作比較，左右顯示部分採用隨機顯示方式，刻意的去隱瞞我們的網站與欲比較的網站，讓使用者沒有先入為主的觀念下，對於我們景點的推薦進行評比。

圖 3.1Blind Test 實驗方式

我們的系統除了呈現所探勘出的景點外，更呈現了景點熱門程度的部分，我們根據我們所探勘出該景點的樣本數，來分析出該點是否為較熱門的景點，較熱門的景點以較大的圖示來表示，利用點擊次數來評估使用者對於該景點的興趣程度，以得知我們的過濾方式所探勘
的樣本，是否能呈現景點的熱門程度。

圖 3.2 模擬熱門景點以 GoogleMap 方式呈現畫面

另外根據我們的景點推薦系統以傳統電子問卷的方式進行調查，在這部分主要是想要獲得該景點的地區性滿意度，如：在地人所知道的有趣景點，是否有被我們的系統探勘出來。

最終使用者滿意度調查

推薦的景點是否滿意
○ 5分 ○ 4分 ○ 3分 ○ 2分 ○ 1分

家鄉景點是否被探勘
○ 5分 ○ 4分 ○ 3分 ○ 2分 ○ 1分

系統流暢度與美觀
○ 5分 ○ 4分 ○ 3分 ○ 2分 ○ 1分

..........

..........

○ 5分 ○ 4分 ○ 3分 ○ 2分 ○ 1分

圖 3.3 模擬使用者意見調查問卷畫面

最後我們將收集各地區著名景點位置資訊與本系統做比較，在這
裡的著名景點，如：陽明山、阿里山等經典景點，來評估使用本系統是否能將傳統的景點探勘出來，判斷我們系統的可信度。
第四章 結論與未來展望

我們有效的利用EXIF裡所記錄的照片資訊以及照片來源網頁所探勘出的相關旅遊資訊，再進一步的使用我們所定義的分析及過濾方式，得到去除雜訊的旅遊景點，最後再使用Pagerank的方式，去除住家、辦公室等個人活動場所，以景點熱門程度的方式來推薦旅遊景點給使用者。我們在抓取照片的同時更擷取了資料所在網頁其他資源，一個旅遊景點對於不同語言、景點名稱的描述，可供使用者查詢到更為通俗的景點名稱，另外使用者對於一個有興趣的旅遊景點便能直接連接至該照片所在網頁，不用再額外的使用搜尋工具、搜尋引擎、論壇網站、知識家等等尋找欲查詢的旅遊景點。最後我們使用我們所設計的實驗方法，能有效的得到自動取得景點相關資訊的準確率，也得夠得到自動取得旅遊景點的準確率及可信度、熱門景點推薦的資訊可信度。

未來我們將可更進一步的將實驗範圍拓展至探勘全世界的旅遊景點，由於現在有越來越多使用者採用自助旅行，我們也可朝推薦旅遊序列領域開始研究，如何有效的推薦不同的個人化旅遊序列，在目前觀光發展的時代，相信是目前蠻重要研究議題。
第五章 相關研究

在 Y.Yang 的論文裡，他們使用 OWL 的方式來收集與建立旅遊資訊，OWL 是 Web Ontology Language 的縮寫，OWL 旨在用於那些需要由應用程式而不是由人類來處理檔案中資訊的情形，也就是說能被機器所了解的語言。OWL 可被用來明確表示辭彙表中術語的含義以及術語間的關係。在表達含義和語義方面，OWL 比 XML、RDF 和 RDF-S 這類語法有更多的表達手段，因此在 Web 上表達電腦可理解內容的能力也比這些語言強，這種 OWL code 主要被設計用來處理資訊的内容，本篇 Paper 利用此特性針對網路上眾多使用此方式的旅遊網站，進行解析並整合資訊。

![圖 5.1 OWL 編輯器](image)

如圖 5.1 他們使用一種 OWL 的編輯器，圖裡的圖形表示其概念、線表示其概念間的關係，讓使用者輕易訂定自我的概念及關係語意。
未來我們可利用此篇 Paper 的方法自動的取得此類方式建置的旅遊資訊，但對於此種語法的網站還是相當的少，所以我們能獲得的資訊相當有限。

A.Popescu 在 2009 年，他們使用使用者所提供的照片與取得的景點名稱做匹配，在使用者所提供的照片部分，它們分析使用者所更改的照片檔名，或者該照片所在網頁所描述的標題與內文，在景點名稱部分，使用 wiki 來取得景點的名稱建立資料庫，另外利用 wiki 的廣大資源來取得其餘十個使用較為頻繁的國外語言，在這裡他們建造了多語言詞彙的地理內容，來增加其 match 的範圍。

從圖 5.2 來看，左邊為該景點個數為含有一百張以內的旅遊照片，右邊為大於一千張旅遊照片的景點個數，由此我們可知，從前的著名景點所包含的個數較多，也廣為人們所拜訪，對於此樣本的收集，可探勘出許多相較之下較不熱門的景點。對於此篇 Paper 的方法，wiki 上不見得所有景點都有收集在內，所介紹的景點可能還是廣為人
們所知的景點，但不見得是最為熱門的景點，所以有下表的差異行為，另外藉由人工方式所更改的暱稱與網頁所描述的資訊所收集的樣本數，可能還有待思考。

在 2009 年 Y.Zheng 這篇以 GPS 軌跡紀錄使用者出外旅遊的活動行為，來挖掘出有趣的景點，首先他們先以所定義的停留點，裡面包含了兩個參數：時間門檻以及距離門檻，來判斷旅遊景點。再使用密度叢集演算法來消除雜亂資料，將所收集的資訊群集分層次，這裡所指的雜亂資料為群集後密度較為稀疏的地方，接著使用 TBHG 來模組化此旅遊序列。

![圖 5.3 TBHG 架構圖](image)

首先先以上述方式收集完資訊，將群集分層次如圖 5.3 右所示，再建立圖以有向連接線的方式連結其景點間相互關係。再以 HITS 推論模組來直接連結使用者和拜訪的景點，在這裡使用了兩個參數：Authority 代表景點的分數，這裡的景點分數為其指向此景點的 Hub 分數的加總分數，分數高表示被許多高頻率旅遊的使用者所拜訪，
Hub 代表使用者的分數，這裡的使用者分數為所指 Authority 分數的總合，分數高表示該使用者拜訪許多高頻率的旅遊景點，接著使用迭代法的方式，不斷地計算其分數，來排名景點與使用者。本篇 Paper 使用 Gps log 的方式，不見得每個人都有該裝置，較難以收集有效地旅遊資訊，另外使用 HITS 的方式，分數較容易造假，因為一個使用者可能旅遊的次數越多其分數越高，但不見得它所去的景點都是有趣的景點。
第六章 參考文獻 (未整理)

一次解讀 Pagerank: http://mmdays.com/2007/06/05/pagerank-explained/

維基百科 Pagerank：http://zh.wikipedia.org/zh-tw/Pagerank

[1] “中華民國交通部觀光局行政資訊系統.”


[5] 計.，照片地理標記（Geotagging Photo）處理流程, 中央研究院，


