ONLINE REPRESENTATION OF TOURISM: AN ANALYSIS OF SEARCH RESULTS FROM A MAJOR SEARCH ENGINE

Zheng Xiang
National Laboratory for Tourism & eCommerce, School of Tourism & Hospitality Management, Temple University

Daniel R. Fesenmaier
National Laboratory for Tourism & eCommerce, School of Tourism & Hospitality Management, Temple University

Martin Yongho Hyun
National Laboratory for Tourism & eCommerce, School of Tourism & Hospitality Management, Temple University

Karl Wober
Institute for Tourism and Leisure Studies, Vienna University of Economics and Business Administration

Online Representation of Tourism: An Analysis of Search Results from a Major Search Engine

Zheng Xiang
Daniel R. Fesenmaier
Martin Yongho Hyun
National Laboratory for Tourism & eCommerce
School of Tourism and Hospitality Management
Temple University

Karl Wöber
Institute for Tourism and Leisure Studies
Vienna University of Economics and Business Administration

ABSTRACT

The goal of this study is to provide a preliminary assessment of the representation of tourism destinations through a major search engine. A three-step analysis was conducted with the focus on assessing: 1) the visibility of tourism-related information regarding 30 tourist destinations within the United States; 2) the visibility of various industry sectors within 3 selected destinations; and, 3) a comparison of domain URLs of search results on a major destination. The findings show that, although there is huge amount of information indexed travelers can only access a tiny fraction of this information. Also, there are a number of dominant players among the Web pages suggested by the search engine. This study provides insights into the challenges the tourism industry is faced with when promoting a destination. It also offers several implications for developing tools and marketing strategies for the tourism industry.

INTRODUCTION

The Internet has become one of the most important information sources for travel related activities (TIA, 2005). While the Internet provides a useful technological and communication platform for tourism organizations to market their products and services to prospective visitors, effectively organizing information over the Web to support travel information search has always been a challenging task (Werthner & Klein, 1999). Recent technological trends on the Internet seem to have escalated these challenges. Particularly, general purpose search engines such as Google and Yahoo! have become increasingly dominant in affecting how people access the cyberspace (Spink et al., 2002; TIA, 2005). It is widely known that these search engines, while providing "global" coverage with millions of search results for even one single query, offer little precision in locating the most relevant information for users. Considering the fact that the tourism industry is comprised of a large number of small- and medium-sized individual operations within one destination, it further worsens the situation for tourism businesses in gaining accessibility and visibility on the Internet, leading to a new type of competition and consumption of their limited resources (Google, 2006; Wöber, 2006).

The goal of this study was to assess the current status of the online representation of the tourism industry. Specifically, it intended to answer the following research question: Given a specific tourist destination, to what extent is it represented through a search engine on the

Internet? By addressing this question, this study offers insights into the challenges the tourism industry is faced with in promoting the destination.

RESEARCH BACKGROUND

The Internet along with the applications used to support navigation and information retrieval can be seen as a communication platform between the consumer and various businesses (Hoffman & Novak, 1996). From the tourism industry's perspective, the task is to effectively organize relevant information and provide the means by which travelers can utilize the information for different purposes. In order to accomplish this goal, it is important to identify the issues and challenges resulting from the market and technological dynamics on the Internet (Werthner & Klein, 1999). As such, it is essential to understand the status of representation of the tourism product on the Internet. Generally speaking, tourism is understood as a "system" which is geographically bounded and consists of both the industry facets and the consumption experience of travel (Mill & Morrison, 2003; Pearce, 1987; Smith, 1988; Woodside & Dubelaar, 2002). Since language mediates communication, the language used to promote tourism and the language used by tourists form the basis for interpreting tourism products and travel experiences (Dann, 1997). Many authors have examined and critiqued the role of language in tourism marketing and promotion in creating, inducing, and reinforcing imageries of tourism products.

Within a technological context, recent research on online travel information search (e.g., Pan & Fesenmaier, 2006; Xiang et al., In press) has found that online tourism marketers use significantly different languages from travelers. That is, a large amount of marketing-oriented content is focused on the selling of travel products, while travelers use more subjective and experiential language to describe their perceptions and expectations when searching for travel products. The gap between the promotion of travel products and travel information search indicates that marketers need to utilize different forms of communication to enable travelers to express their need for information that is framed within their personal context. From the system design point of view, Gretzel and Fesenmaier (2002) argue that the existing approaches in tourism information systems, while generally relying on numeric data, cannot capture the holistic experiential aspects of travel. It is argued that narrative logic should be incorporated into online systems to assist travelers to make sense of the world. Along the same line, Xiang and Fesenmaier (2006) conclude that the use of interface metaphors can help bridge the gap between the rich and holistic experience of travel and the lack of means to convey it in an online system.

It seems that from a marketing viewpoint the challenges for information provision in tourism in large part arise from the role the Internet and related technological applications play in mediating the communication between the traveler and the industry. While the existing emphasis on the language representation of tourism on the Internet has provided important implications for designing and improving online marketing efforts, little has been done in terms of the extent to which the Internet mediates the visibility and accessibility of tourism products before the actual interaction can be established. Recently, Wöber (2006) conducted an analysis of the visibility of destination marketing organizations and individual hotel operations in Europe among six popular search engines. The findings show that many of these tourism businesses enjoy very low rankings among the search results, which makes it extremely difficult for online travelers to directly access individual tourism operations through these search engines. Thus, given the growing importance of search engines in online information search and transaction, it appears that very little is known about the representation of the tourism industry across various destinations and industry sectors through these technological tools.

METHODOLOGY

To provide a comprehensive understanding of the online representation of tourism, a three-step content analysis approach was employed using search results retrieved from Google. These three steps include: 1) mapping a pool of sampled tourist destinations in the United States, ranging from internationally reputed ones to regional to less known ones, by showing their visibility on Google; 2) mapping the industry sectors within tourist destinations by showing their visibility on Google based upon search results retrieved using 20 predefined queries; and, 3) analyzing the domain names of retrieved search results using the 20 queries used in the previous step for a single destination to show the important information providers in the industry.

For each of the three tasks, a Web mining program written in Perl programming language was developed and used to retrieve the results from Google. The results were parsed by the program in real time and the following indices were saved into a flat text file: 1) the total number of Web pages Google has indexed (i.e., based on the snippets such as "Results 1 - 10 of about *143,000,000* for tourism" shown on Google's result page), 2) the total number of search results that actually can be displayed, as well as 3) the URLs, rankings, titles, and abstracts for each of the search results. The total number of indexed pages is used to represent the possible size of the domain for the specific query, and the total number of search results displayed is used to represent the possible size of the domain for the specific query that is made visible and accessible by the search engine for the user. As a well-known rule of thumb, Google makes no more than 1,000 search results accessible to the user. An index called "visibility ratio" was calculated by dividing the number of results presented by the total number of Web pages indexed by Google.

In Step 1, 30 U. S. cities were selected to represent the population of tourist destinations in the United States. Among these 30 destinations, the first group of 10 represents the most popular destinations with international reputations (including New York, etc.); the second group of 10 cities represents destinations with mainly national fame; and, the third group of 10 cities represents those destinations with a regional influence. The cities among the first two categories, i.e., major destinations and secondary destinations, were obtained through an online list of consumer-voted most popular destinations in the United States (www.virtualtourist.com). Cities among the third category, i.e., minor destinations, were randomly picked based upon a complete list of destination marketing organizations compiled by the authors. As can be seen in the result section, these destinations have a diversity of attractions and are geographically distributed across the United States. The queries used for these destinations were a combination of the destination name and the keyword "tourism".

In Step 2, three destinations were chosen with one from each of the categories in Step 1. They are Chicago, IL, Charlotte, NC, and Elkhart, IN. Twenty keywords were selected in combination with the destination name to form 20 queries. The selection of these keywords was guided by both the classification schemes from the industry and the actual queries used by travelers. Websites of destination marketing organizations (e.g., www.choosechicago.com) were used as sources to identify the keywords used to describe the types of information provided on the websites (e.g., "accommodation", "activities", "attractions", etc). The results of a study of a European-based search engine which contains about 180,000 entries of users' query logs (Wöber, 2006) were used to identify keywords used by travelers when they were seeking destination-related information. The resulting 20 queries (i.e., the combination of destination name and each of the 20 keywords) represent the hierarchical structure of decision making during travel

planning as well as the most important business facets in the industry (Smith, 1988; Woodside & Dubelaar, 2002).

In Step 3, the URLs retrieved from the search results for Chicago in Step 2 were analyzed in order to gain an understanding of the tourism related information represented by the search results. According to the information retrieval literature (e.g., Spink et al., 2002), most search engine users (>85%) do not go beyond the third page to view search results. As such, the first 30 URLs which constitute the first 3 pages of search results were extracted to represent the search results that are most likely to be viewed by travelers. In addition, another 20 URLs were extracted with 10 ranked in the middle of the overall search results and the other 10 at the bottom of the entire search result set. This resulted in 50 URLs for each query and a total of 1,000 URLs. The results for the 20 queries were then labeled into 5 groups to represent the first page, second page, third page, and the middle page, and the bottom 10. The URLs were further parsed by extracting the part containing the top-level domains. That is, a URL such as "http://www.somedomain.com/somedirectory/somepage" results in a string like "http://www.somedomain.com". Last, frequencies of each unique domain URLs were calculated for each of the 5 categories.

FINDINGS

Table 1 shows the results of the first set of analyses. As can be seen, Google indexes a large number of Web pages that are related to the tourism domain. Overall, there are 65,961,000 Web pages that are related to these 30 tourist destinations based on the specific query. However, most of the indexed pages are not visible to search engine users. That is, only a very small fraction of indexed Web pages are shown as search results (mean=696 Web pages) that are accessible to a user, resulting an overall visibility ratio of 0.032%. Considering that most of search engines users only view the first three search result pages, the actual visibility ratio is much lower. Understandably, the more famous a destination is, the more Web pages are indexed by Google. However, it is somewhat paradoxical that, due to the larger numbers of Web pages indexed by Google for more famous destinations, their actual visibility ratios are lower than the less famous destinations (0.015%, 0.050%, and 0.106% for the major, secondary, and minor destinations, respectively).

Table 2 summarizes the results of the Step 2 analyses. Overall, Google generated a huge amount of information as reflected by the total numbers of indexed pages for the three destinations. Among the 20 keywords used to query Google for the three destinations, the word "information" generated the largest number of indexed Web pages for both Chicago and Charlotte. Also, words such as "guide", "sports", and "park" generated relatively large number of indexed pages for these three destinations. One possible explanation is that these words are relatively generic terms, which are not necessarily tied to tourism. As such, the domains these keywords represent are larger than the domains represented by tourism specific keywords. A glimpse of the results also confirms the findings in Step 1 analysis in that the visibility ratios diminishes with the increase of the popularity of the destination, on almost all keywords. It is interesting to note that for some queries (e.g., "Chicago AND dining", "Chicago AND sports", "Charlotte AND accommodation", "Charlotte AND dining", and "Charlotte AND restaurants"), Google only presents a very small set of search results (in the 300s and lower 400s).

Table 1 Mapping the Representation of Destinations on Google (query: Destination Name plus "tourism")

	Destination	Google Results: Total Indexed	Google Results: Results Presented	Google Results: Visibility Ratio
1	New York City	23,100,000	816	0.004%
2	Las Vegas	3,650,000	733	0.020%
3	Chicago	3,630,000	526	0.015%
4	Orlando	3,230,000	572	0.018%
5	Los Angeles	3,100,000	712	0.023%
6	San Francisco	3,080,000	720	0.023%
7	Atlantic City	1,940,000	726	0.037%
8	Philadelphia	1,760,000	716	0.041%
9	Houston	1,730,000	805	0.047%
10	New Orleans	1,500,000	832	0.056%
	Subtotal	*46,720,000*	*7,158*	*0.015%*
11	Phoenix, AZ	2,640,000	595	0.023%
12	Charlotte, NC	1,140,000	508	0.045%
13	Baltimore, MD	1,130,000	593	0.053%
14	Fort Worth, TX	1,120,000	526	0.047%
15	San Jose, CA	1,110,000	909	0.082%
16	Indianapolis, IN	1,080,000	552	0.051%
17	Columbus, OH	1,080,000	671	0.062%
18	Detroit, MI	1,010,000	641	0.064%
19	Jacksonville, FL	907,000	475	0.052%
20	Memphis, TN	653,000	426	0.065%
	Subtotal	*11,870,000*	*5,896*	*0.050%*
21	Myrtle Beach, SC	1,180,000	958	0.081%
22	Lancaster, PA	1,130,000	797	0.071%
23	Newport, WA	1,130,000	754	0.067%
24	Lincoln, NE	1,070,000	979	0.092%
25	Norfolk, VA	887,000	685	0.077%
26	Elkhart, IN	688,000	728	0.106%
27	Pueblo, CO	524,000	646	0.123%
28	Evanston, IL	276,000	559	0.203%
29	Americus, GA	244,000	845	0.346%
30	Bradenton, FL	242,000	874	0.361%
	Subtotal	*7,371,000*	*7,825*	*0.106%*
	Total	65,961,000	20,879	0.032%

Table 2 Mapping Industry Sectors for Three Destinations (Query: destination name plus keyword)

	Chicago, IL			Charlotte, NC			Elkhart, IN		
Keyword	Pages Indexed	Results Presented	Visibility Ratio	Pages Indexed	Results Presented	Visibility Ratio	Pages Indexed	Results Presented	Visibility Ratio
accommodation	1,270,000	892	0.070%	529,000	***400***	0.076%	84,100	514	0.611%
activities	44,600,000	532	0.001%	1,240,000	650	0.052%	726,000	828	0.114%
attractions	7,440,000	706	0.009%	1,170,000	757	0.065%	281,000	736	0.262%
bar	51,000,000	700	0.001%	1,270,000	790	0.062%	887,000	916	0.103%
cultural	46,900,000	940	0.002%	1,100,000	752	0.068%	229,000	852	0.372%
dining	13,600,000	***387***	0.003%	1,230,000	***418***	0.034%	363,000	736	0.203%
event	81,500,000	862	0.001%	1,330,000	938	0.071%	865,000	844	0.098%
festivals	2,440,000	523	0.021%	1,190,000	817	0.069%	296,000	826	0.279%
guide	102,000,000	855	0.001%	1,140,000	775	0.068%	1,210,000	642	0.053%
hotel	66,600,000	832	0.001%	1,200,000	***373***	0.031%	979,000	910	0.093%
information	246,000,000	832	0.000%	1,630,000	800	0.049%	1,170,000	982	0.084%
motel	1,380,000	916	0.066%	1,060,000	475	0.045%	556,000	830	0.149%
nightlife	1,990,000	838	0.042%	582,000	586	0.101%	32,400	429	1.324%
park	117,000,000	910	0.001%	1,340,000	644	0.048%	1,190,000	982	0.083%
restaurants	28,500,000	800	0.003%	1,170,000	***371***	0.032%	515,000	515	0.100%
shopping	73,500,000	923	0.001%	1,330,000	727	0.055%	950,000	660	0.069%
sports	94,900,000	***363***	0.000%	1,380,000	830	0.060%	1,220,000	784	0.064%
theater	34,700,000	880	0.003%	1,230,000	640	0.052%	530,000	979	0.185%
things to do	42,600,000	724	0.002%	1,430,000	***373***	0.026%	448,000	698	0.156%
transportation	20,400,000	761	0.004%	1,210,000	695	0.057%	1,110,000	915	0.082%
Total	1,078,320,000	15,176	0.001%	23,761,000	12,811	0.054%	13,641,500	15,578	0.114%

Figure 1 presents the results of the third set of analyses. The plots show the cumulative percentages of unique domain names identified among 200 URLs of search results for different sections of result pages on Google (i.e., the first, second, third pages, the page in the middle, and the last 10 results, respectively). In addition to these result sections, a baseline was drawn to show the cumulative percentage over 200 search results with the assumption that all the search results have unique domain names. As shown in the graph, the first page of search results contains the highest number of duplicate domain names. That is, with 25 unique domain names, the cumulative frequency reaches approximately 60% of the total frequency (200), which is almost 5 times as high as the cumulative frequency (25/200=12.5%) in the baseline. The level of duplication diminishes with the "depth" of access (i.e., from the first to second to third page and to "deeper" pages). However, it seems that there is a relatively high level of duplication among the bottom 10 search results. Referring to the data, it can be seen that most of domain duplicates are 1) portals and information aggregators such as chicagotraveler.com, citysearch.com, Yahoo! Travel, and Chicago Tribune Online; 2) destination marketing organizations' sites such as choosechicago.com; and, 3) government site such as cityofchicago.org.

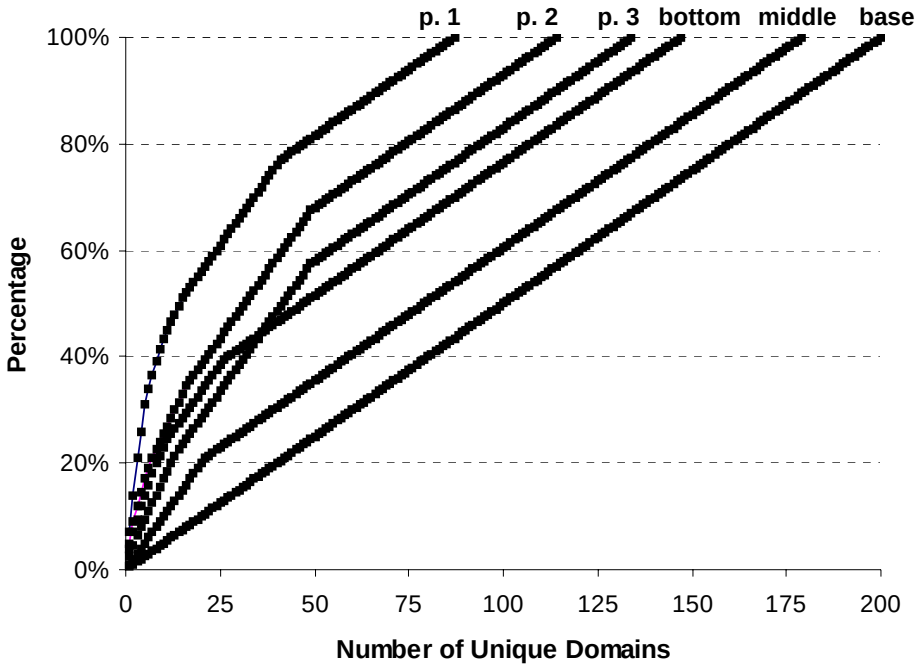

Figure 1 Distribution of Unique Domain Names among Search Results on Different Sections of Google Result Pages

DISCUSSION AND CONCLUSIONS

This study examined the extent to which the tourism industry in represented on the Internet through a major search engine. The results show that while the search engine claims that it has indexed a very large number of Web pages that are related to tourism, the actual size of the tourism domain that is made visible and accessible to travelers is only a tiny fraction of total number of potentially relevant Web pages. This demonstrates the so-called "thin" interface effect in that the currently existing technologies are faced with challenges in representing a complex domain like tourism. In addition, these analyses show that due to the limitations inherently resulting from the way the search engine has been designed a huge amount of potentially useful information has been filtered out. This is consistent among various tourist destinations and across different sectors with the tourism industry. The low visibility of tourism Websites in general purpose search engines indicates that the chances are slim for online travelers to have direct interactions with many tourism enterprises and organizations. This study also shows that there are a number of dominant players among the Web pages that are indexed and presented by the search engine. The salient presence of portal websites and information aggregators indicates that a general purpose search engine like Google is ineffective in terms of organizing and providing tourism related information in a meaningful way.

From the search engine's point of view, it can be argued that the queries used in this study are simplistic in that they only contain one or two keywords. Essentially, more elaborate queries can be used to locate more specific information through search engine, as long as a match can be established regardless of what matching algorithms are being used. However, from the user's perspective, travelers may not have the capabilities to come up with more sophisticated queries, which has been shown in literature on human-computer interaction (Furnas

et al., 1987) and recent research specifically on search engines (Jansen & Molina, 2006). Therefore, even though in theory more specific tourism related information can be made visible to travelers, in reality it is extremely difficult for travelers to actually access the "deeper" levels of the tourism domain.

This study offers substantial insights into the challenges the tourism industry is faced with regarding their strategies to gain access to prospective visitors. Specifically, it indicates that there is a need to identify new solutions to improve the visibility of the tourism industry on the Internet as well as innovative ways to represent tourism information to travelers in search of experiential encounters. It seems that the recent development of domain specific search engines, destination portals, and tourism recommender systems are essential (Fesenmaier, Wöber, & Werthner, 2006; TIA, 2006; Wöber, 2006).

However, research is needed to identify viable and meaningful ways to meet many of the challenges resulting from the market and technological dynamics. Following from this study, there are a number of directions for future research. First, a more in-depth analysis of search results from Google should be conducted by categorizing the search results based upon the nature of the information in supporting travel planning. Second, analyses need to be conducted of other major search engines such as Yahoo! and Ask to understand the commonalities and differences between these technologies in representing the tourism domain. Third, research should be conducted to gain a better understanding of how travelers actually respond to the representation of tourism on search engines.

REFERENCES

Dann, G. M. S. (1997). *The Language of Tourism: A Sociolinguistic Perspective*. Wallingford, UK: CAB International.

Fesenmaier, D. R., Wöber, K., & Werthner, H. (2006). Introduction: recommendation systems in tourism. In D. R. Fesenmaier, K. Wöber & H. Werthner (Eds.), *Destination Recommendation Systems: Behavioral Foundations and Applications*. Wallingford, UK: CABI.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM, 30*, 964-971.

Google. (2006). Seattle's Convention and Visitors Bureau found 30% ROI with Google AdWords. Retrieved December 15, 2006, from http://www.google.com/ads/scvb.html

Gretzel, U., & Fesenmaier, D. R. (2002). Building narrative logic into tourism information systems *IEEE Intelligent Systems* (November/December), 59-61.

Hoffman, D. L., & Novak, T. P. (1996). Marketing in hypermedia computer-mediated environments: Conceptual foundations. *Journal of Marketing, 60*, 50-68.

Jansen, B. J., & Molina, P. R. (2006). The effectiveness of Web search engines for retrieving relevant ecommerce links. *Information Processing and Management, 42*, 1075-1098.

Mill, R. C., & Morrison, A. D. (2003). *The Tourism System: An Introductory Text* (4th ed.). Dubuque, IA: Kendall/Hunt Publishing Company.

Pan, B., & Fesenmaier, D. R. (2006). Online information search: vacation planning process. *Annals of Tourism Research, 33*(3), 809-832.

Pearce, D. (1987). *Tourism Today: A Geographical Analysis*. New York: Longman Scientific & Technical.

Smith, S. L. J. (1988). Defining tourism: a supply-side view. *Annals of Tourism Research, 15*, 179-190.
Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *IEEE Computer, 35*(3), 107-109.
TIA. (2005). *Travelers' Use of the Internet*. Washington, DC: Travel Industry Association of America.
TIA. (2006). *Distribution Strategy: Emerging Models for Destination Portals*. Washington, DC: Travel Industry Association of America.
Werthner, H., & Klein, S. (1999). *Information Technology and Tourism: A Challenging Relationship*. Vienna: Springer.
Wöber, K. (2006). Domain specific search engines. In D. R. Fesenmaier, K. Wöber & H. Werthner (Eds.), *Destination Recommendation Systems: Behavioral Foundations and Applications*. Wallingford, UK: CABI.
Woodside, A. G., & Dubelaar, C. (2002). A general theory of tourism consumption systems: A conceptual framework and an empirical exploration. *Journal of Travel Research, 41,* 120-132.
Xiang, Z., & Fesenmaier, D. R. (2006). Interface metaphors on travel-related Websites. In D. R. Fesenmaier, K. Wöber & H. Werthner (Eds.), *Destination Recommendation Systems: Behavioral Foundations and Applications* (pp. 180-189). London: CAB International.
Xiang, Z., Kim, S.-E., Hu, C., & Fesenmaier, D. R. (In press). Language representation of restaurants: implications for developing online recommendation systems. *International Journal of Hospitality Management*.

Contact Information:

Zheng (Phil) Xiang
School of Tourism and Hospitality Management
Temple University
1700 N. Broad Street, Suite 316E
Philadelphia, PA 19122
Tel: 215-204-1941
Fax: 215-204-8705
Email: philxz@temple.edu