Proceedings of I-KNOW '05
Graz, Austria, June 29 - July 1, 2005

Where Do You Want to Go Today? A Media Analysis of Global Tourism Destinations

Astrid Dickinger[†], **Arno Scharl**[*]
(Vienna University of Economics & Business Administration, Austria
[†]Institute for Tourism and Leisure Studies, [*]Information Systems Department
{astrid.dickinger, arno.scharl}@wu-wien.ac.at)

Albert Weichselbraun
(University of Western Australia, School of Computer
Science and Software Engineering, Australia
weichselbraun@ecoresearch.net)

Abstract: Destinations are at the heart of travel decisions, and destination image has a significant influence on tourists' decision-making. Many travelers acquire information via the Internet, which offers abundant information and an increasing number of tourism-related services. The impact of media coverage on destination image has attracted research attention and became particularly evident during the 2003 outbreak of SARS, the Severe Acute Respiratory Syndrome. Building upon previous research, this paper analyzes the prevalence of tourism destinations among 162 international media sites. Measuring *term frequency* investigates the attention a destination receives – from a general and, after *contextual filtering*, from a tourism perspective. Calculating the *semantic orientation* estimates positive or negative media influences on destination image at a given point in time. By detecting associations with country names, *keyword analysis* reveals the countries' public profile, and the impact of events on media coverage.

Keywords: Tourism Destinations, Media Monitoring, Lexical Analysis, Semantic Orientation
Categories: H.3.1 Content Analysis and Indexing, H.3.3 Information Search and Retrieval, K.4.4 Electronic Commerce

1 Introduction

Trends in global tourism have shifted remarkably over the last decade. Information technology supports the increased sophistication of travelers [Chen and Sheldon 1997], who seek greater variety in their travel arrangements and expect personalized services that meet their unique needs [Sheldon 1993]. Previously, travelers received information about destinations through books, brochures, promotional videos, word-of-mouth, travel agents or tourist offices. Radical changes in acquiring destination information in recent years have increased the challenge for destination managers to promote a complex product in a highly competitive industry. To meet this challenge, many organizations focus on image studies, marketing strategies, conversion studies and advertising research [Dore and Crouch 2003], but neglect publicity through unplanned and incidental news and media coverage.

In the past, the process of collecting data was time consuming, expensive, and often resulted in outdated and incomplete information. Nowadays detailed destination profiles are readily available online, allowing for inexpensive, fast and topical re-

search. Still considered an emerging medium, the Internet is already an important means of gathering trusted destination information [Cai et al 2004]. Yet travelers often find it difficult to identify high-quality sources, while destination marketers struggle to differentiate and promote their offerings in the flood of information.

Surveys have shown that available information influences destination choice, tourist satisfaction, purchase decision behavior and the likelihood of repeat visits [Perdue 1985, Guy et al 1987, Cai et al 2004]. Travelers have to evaluate an intangible product at the time of purchase evaluation, as destinations cannot be physically presented at the point of sale. Thus destination marketers have to provide detailed descriptions, photos and videos to influence the consumer purchase decisions, including the intended length of stay and level of expenditure [Fesenmaier 1994].

This paper investigates the prevalence and image of destinations in online news media as of April 2005. It compares general and tourism-specific coverage, as the depiction of a country as a tourism destination does not necessarily correlate with the country's overall profile. Such a comparison promises new insights into the structure of destination coverage, and serves as an indicator for the relative importance of tourism for the economy of individual destinations.

2 Methodology

Tourism is among the leading applications of electronic commerce technology, with a high percentage of actors maintaining independent Web sites. General and domain-specific search engines based upon increasingly effective Web crawler technology are crucial for locating destination information online. Among the main challenges with regard to achieving high search engine rankings is the often poor visibility of destination marketing organizations, inadequate content, the favoring of well-known destinations, and generic search strategies of many potential visitors. To address these challenges, [Delgado and Bowen 2004] call for destination portals that retrieve, match and deliver destination information based on advanced Web crawlers and lexical-statistical processing of semantic information.

The project presented in this paper lays the foundation for such a portal by aggregating fragmented tourism information [Maedche and Staab 2002]. The webLyzard crawling agent (www.weblyzard.com) mirrors a selection of international news media sites from the *Kidon.com*, *ABYZNewsLinks.com* and *NewsLink.org* directories in weekly intervals. The sample comprises 162 sites from seven English-speaking countries: United States (62), United Kingdom (42), Canada (17), Australia (19), South Africa (9), New Zealand (8) and Ireland (5). The crawling agent considers both visible (e.g. raw text including headings, menus and link descriptors) and invisible text (e.g. embedded markup tags and scripting elements). Excluding graphics and multimedia files, the crawling agent follows a site's hierarchical structure until amassing 50 megabytes of textual data. The size restriction helps compare systems of heterogeneous size, and manage storage capacity. Documents of lower hierarchical levels that do not belong to the primary user interface (e.g. newspaper archives) fail to reflect typical usage patterns.

The parsing component splits the retrieved textual chain into sites, documents and sentences. The hierarchical, XML-encoded output file thus preserves the original site structure. The system then identifies and removes redundant copies of news headlines

and non-contextual navigational elements, whose appearance on multiple pages would otherwise distort frequency counts – particularly when using contextual filtering as described in [Section 3.2].

3 Results

Media coverage on destinations tends to fluctuate heavily due to political events, scientific discoveries, military operations or natural disasters. To measure the extent of coverage on particular destinations, a case-insensitive pattern matching algorithm processed regular expression queries (formalisms describing a set of strings without enumerating its elements) based on the names of countries, capitals and large cities with a population of more than a million. In the case of Austria, for example, the system queried the tokenized output for the following: *^austria$*, *^graz$*, *^innsbruck$*, *^linz$*, *^salzburg$* and *^vienna$*.

The lack of local context limits the explanatory power of word frequency data [McEnery and Wilson 1996, Biber et al 1998]. Only measuring the number of occurrences neglects the author's attitude, an important aspect of the human language. Assuming that text segments reflect local coherence, author attitude can be inferred from the distance between a target term and sentiment words from a tagged dictionary [Scharl et al 2004, Scharl 2004]. The current dictionary uses 4,400 positive and negative sentiment words from Harvard's General Inquirer [Stone 1997]. Reverse lemmatization added about 3,000 terms by considering plurals, gerund forms and other syntactical variations (e.g. manipulate → manipulates, manipulating, manipulated).

3.1 Media Attention and Destination Image

[Tab. 1] shows that the most frequently mentioned country was the United States, followed by Canada, Australia, the United Kingdom and Iraq. Given the Anglo-American news media sample and recent geopolitical events, this sequence is hardly surprising. Countries rarely mentioned were Mayotte, Saint Pierre and Miquelon, Saint Lucia, Svalbard, Guadeloupe and the Northern Mariana Islands. In terms of positive semantic orientation, Palau takes the lead, followed by San Marino, Bahrain, Saint Lucia and Niue. Negative coverage concentrated on Iraq, Vietnam, Angola, Somalia and Equatorial Guinea.

Frequent Countries	Freq	SO
USA	144,079	.121
Canada	63,548	.112
Australia	43,052	.133
UK	40,636	.114
Iraq	35,231	-.197
Ireland	23,869	.091
France	19,416	.110
China	18,167	.093
New Zealand	11,970	.114
Italy	11,002	.094

Rare Countries	Freq	SO
Faroe Islands	44	.176
Pitcarin	40	.016
Tokelau	30	.075
Martinique	30	-.070
N Mariana Islands	22	.161
Guadeloupe	22	-.055
Svalbard	20	.072
Saint Lucia	19	.270
Saint Pierre & Miquelon	15	.123
Mayotte	13	.164

Positive Coverage	SO	Freq
Palau	.325	181
San Marino	.282	93
Bahrain	.276	556
Saint Lucia	.270	19
Niue	.252	75
Saint Kitts & Nevis	.245	96
New Caledonia	.245	73
Micronesia	.229	58
Gabon	.195	87
Belize	.195	218

Negative Coverage	SO	Freq
Eritrea	-.105	177
Kuwait	-.107	753
Congo	-.110	1,542
Kyrgyzstan	-.123	768
Sudan	-.143	2,284
Equatorial Guinea	-.144	201
Somalia	-.149	566
Angola	-.156	843
Vietnam	-.176	3,241
Iraq	-.197	35,231

Table 1: Country rankings by frequency and semantic orientation

Co-occurrence analysis [Roussinov and Zhao 2004] sheds further light on the rankings in [Tab.1], assuming that semantically related terms regularly appear in the same text segments. This research used the *log likelihood algorithm* [Dunning 1994] to identify topics associated with the highest-/lowest-ranking countries in terms of semantic orientation. Limiting the consideration set of co-occurring terms to nouns by means of *part-of-speech tagging* reduced memory consumption and improved both the throughput and the quality of results (part-of-speech tagging analyzes textual corpora to distinguish nouns from articles, verbs, adjectives, etc. [Abney 1996]).

[Tab. 2] illustrates that only some of the identified associations relate to travel and tourism. Other keywords describe TV shows (Palau Survivor), sports events (Bahrain Formula 1 Grand Prix), economic indicators (Palau, San Marino), political processes (San Marino, Bahrain), or military conflicts (Vietnam, Iraq). While even keywords not directly related to tourism are relevant for tourism research because they indicate processes with direct or indirect impact on destination image, the following [Section 3.2] introduces a contextual filtering component that enables analysts to target topics more precisely.

Country	SO	Top 10 Keywords (Significance)
Palau	.325	ulong (38987), koror (37967), bobby jon (28156), ibrehem (26847), stephenie (26276), survivor palau (25539), janu (10489), immunity challenge (7646), preferred stock (6713), jeff probst (6016)
San Marino	.282	parliamentary election (171365), presidential election (105273), legislative election (80144), securities (12767), stock (10796), swap (9354), special-purpose entity (9027), legislative (8750), stopping curve (8094), state tax-free (8094)
Bahrain	.276	schumacher (63484), alonso (62221), ferrari (46198), parliamentary election (27730), renault (20414), barrichello (20394), trulli (18128), presidential election (16989), israeli (15460), prix (14891)
Angola	-.156	marburg virus (27089), luanda (24510), uige (20396), mishawaka (15738), tricentral (15379), decatur (15229), congo (15203), elkhart (15144), vincennes (14372), ebola-like (14139)
Vietnam	-.176	iraq (19858), war (17465), uq wire (14421), kerry (9720), hanoi (6069), shields (6068), vietnam war (5391), discusses (5099), zaoui (4884), bird flu (4666)
Iraq	-.197	war (62321), us (37146), baghdad (30703), saddam (26917), troops (23437), blair (14994), weapons (14640), hussein (13588), soldiers (13582), occupation (10736)

Table 2: Keywords for countries with very positive or negative coverage

3.2 Contextual Filtering

The results presented in the preceding section do not consider whether a news article mentions destinations in a tourism context. The problem can be addressed by *contextual filtering*, which confines the computation of term frequency to occurrences in documents that contain tourism-relevant terms. This research used a contextual filter based on the following set of regular expressions (in alphabetical order):

- accommodations?
- backpack(ing|ers?)
- bed (and|&) breakfasts?
- camping
- destination (images?|information|marketing|positioning)
- holiday(s|ing|makers?)?
- honeymoon(er)?s?
- hos?tels?
- hospitality
- motels?
- national parks?
- nature(|-)?(trails?|parks?)
- sight(|-)?see(ing|rs?)?
- souvenirs?
- tour (guides?|operators?)
- touris(m|ts?)
- travel(s|l?ing|l?ers?)?
- vacation(s|ing)?

Comparing the general corpus with the tourism corpus allows a detailed investigation of tourism coverage and its relative importance for a country. A destination might receive negative coverage due to current events, for example, but still rank high as a tourism destination. Establishing a tourism context also increases the validity of results. CASABLANCA, for example, refers to both Morocco's capital and the famous movie starring Humphrey Bogart and Ingrid Bergman.

3.3 Proportion of Tourism Coverage by Destination

In a tourism context, news articles most frequently refer to the United States, Canada, Australia, Ireland and the United Kingdom. As previously mentioned, this ranking reflects that only Anglo-American news media were sampled for this research.

Of particular interest is the proportion of tourism coverage compared to total media coverage. This ratio hints at the importance of tourism for a destination's economy. Data from the World Tourism Organization such as tourism arrivals and tourism receipts support this assumption, which are shown in [Tab. 3] on a per capita basis.

The Polynesian Islands including Niue, the Cook Islands and French Polynesia, for example, are among the world's prime tourism spots. Thus it does not surprise that nearly 80% of Niue media coverage relates to tourism, followed by Dominica and the Cook Islands (77%), the Maldives (73%), the Cayman Islands (72%), Belize (71%), the Northern Mariana Islands (68%), Martinique (67%), French Polynesia (66%) and the Netherlands Antilles (66%).

On the other end of the scale, there is little tourism coverage on Djibouti (12%), Kiribati (13%), Serbia and Montenegro (13%), Kyrgyzstan (16%), and French Guiana (16%). Reasons for the dominance of information not related to tourism include news alerts on elections in Djibouti, negotiations with the European Union in Serbia and Montenegro, the Taiwanese president's upcoming visit to Kiribati, the volatile political situation and foreign military use of air bases in Kyrgyzstan, and a planned Russian space center at the Kourou cosmodrome in French Guiana.

Max Tourism Coverage	**Freq**	**Freq-T**	**SO**	**SO-T**	**Freq-%**	**ApC**	**RpC**
Niue	75	59	.252	.288	**78.7**	.93	928
Dominica	93	72	.168	.171	**77.4**	1.05	736
Cook Islands	74	57	.037	.061	**77.0**	2.03	1,198
Maldives	303	221	.066	.065	**72.9**	1.66	937
Cayman Islands	191	138	.129	.104	**72.3**	6.82	13,572
Belize	218	154	.195	.153	**70.6**	.81	487
N Mariana Islands	22	15	.161	.017	**68.2**	5.77	8,370
Martinique	30	20	-.070	-.051	**66.7**	1.04	570
French Polynesia	98	65	.190	.253	**66.3**	.80	1,224
Netherlands Antilles	65	43	.070	.128	**66.2**	1.24	3,878
Min Tourism Coverage	**Freq**	**Freq-T**	**SO**	**SO-T**	**Freq-%**	**ApC**	**RpC**
Suriname	68	13	.089	.113	**19.1**	.13	32
Cóte d'Ivoire	48	9	.053	-.018	**18.8**	.01	3
Bahrain	556	96	.276	.203	**17.3**	.04	929
Moldova	209	36	.160	.340	**17.2**	.00	12
Togo	305	52	.080	-.014	**17.0**	.01	2
French Guiana	212	34	-.018	-.061	**16.0**	.34	235
Kyrgyzstan	768	119	-.123	-.205	**15.5**	.01	5
Serbia & Montenegro	1835	244	-.021	-.058	**13.3**	.04	7
Kiribati	70	9	.064	.068	**12.9**	.05	30
Djibouti	82	10	.030	-.042	**12.2**	.04	9

Table 3: Country ranking by tourism coverage (in percent of total coverage)

Variables: Frequency and semantic orientation for the general (Freq, SO) and the tourism corpus (Freq-T, SO-T) as of April 2005, percentage of tourism coverage relative to total coverage (Freq-%), tourism arrivals per capita (ApC), international tourism receipts in US-$ per capita (RpC). The economic indicators are based on the most current data available from the *World Tourism Organization* (www.world-tourism.org).

Planned marketing activities cause or at least influence positive coverage of a country in a tourism context. Negative coverage, by contrast, is often unpredictable and may relate to wars and volatile political situations, food shortages or natural disasters – all representing potential dangers for tourists. Several cyclones that swept through the Cook Islands and left a trail of destruction, for example, lowered the semantic orientation values for this group of islands, normally a favored destination. The low semantic orientation for the Caribbean island of Martinique represents an outlier due to reports on the 2004 crisis in neighboring Haiti and the 1961 death of Frantz Fanon, a West Indian psychoanalyst and social philosopher born in Martinique.

4 Conclusion

Media coverage has a strong impact on the image of tourism destinations. The automated analysis of tourism-related media coverage introduced in this paper helps investigate this impact by revealing the public profile of particular destinations and the impact of current events. Computing the semantic orientation adds an important aspect of the human language, as the frequency of destination references often proves less significant than the attitude conveyed in these references (negative ↔ positive, weak ↔ strong, passive ↔ active, etc.).

Future research will focus on semantic disambiguation and advanced knowledge representations. Automatic knowledge processing requires non-ambiguous terminology, but many tourism terms such as ECOTOURISM or HERITAGE TOURISM leave room for diverging and often conflicting interpretations. Creating shared meaning is the main motivation for building tourism ontologies [Maedche and Staab 2002, Fensel et al 2003]. Future research will integrate ontology knowledge to disambiguate content and facilitate destination tracking. Keyword analysis across hierarchical layers to distinguish between synonyms/antonyms and hyponyms/hypernyms will help to continually extend and revise domain-specific tourism ontologies.

References

[Abney 1996] Abney, S.: "Tagging and Partial Parsing", Corpus-Based Methods in Language and Speech, K. Church, S. Young and G. Bloothooft (Eds.) Kluwer Academic Publishers, Dordrecht, (1996), 118-136.

[Biber et al 1998] Biber, D., Conrad, S. and Reppen, R.: "Corpus Linguistics - Investigating Language Structure and Use", Cambridge University Press, Cambridge (1998).

[Cai et al 2004] Cai, L. A., Feng, R. and Breiter, D.: "Tourist Purchase Decision Involvement and Information Preferences", Journal of Vacation Marketing, 10, 2 (2004), 138-148.

[Chen and Sheldon 1997] Chen, H.-M. and Sheldon, P. J.: "Destination Information Systems: Design Issues and Directions", Journal of Management Information Systems, 14, 2 (1997), 151-176.

[Delgado and Bowen 2004] Delgado, J. A. and Bowen, M.: "DestinationFinder: A Travel Focused Search Engine, Portal and Recommender System for the DMO", Information and Communication Technologies in Tourism (Enter-2004), Springer, Cairo, Egypt, (2004).

[Dore and Crouch 2003] Dore, L. and Crouch, G. I.: "Promoting Destinations: An Exploratory Study of Publicity Programmes Used by National Tourism Organisations", Journal of Vacation Marketing, 9, 2 (2003), 137-151.

[Dunning 1994] Dunning, T.: "Accurate Methods for the Statistics of Surprise and Coincidence", Computational Linguistics, 19, 1 (1994), 61-74.

[Fensel et al 2003] Fensel, D., Wahlster, W., Lieberman, H. and Hendler, J.: "Spinning the Semantic Web - Bringing the World Wide Web to Its Full Potential", MIT Press, Cambridge (2003).

[Fesenmaier 1994] Fesenmaier, D. R.: "Traveller Use of Visitor Information Centers: Implication for Development in Illinois", Journal of Travel Research, 33, 1 (1994), 44-50.

[Guy et al 1987] Guy, B. S., Curtis, W. W. and Crotts, J. C.: "Environmental Learning of First-time Travelers", Annals of Tourism Research, 17, (1987), 419-431.

[Maedche and Staab 2002] Maedche, A. and Staab, S.: "Applying Semantic Web Technologies for Tourism Information Systems", 9th International Conference for Information and Communication Technologies in Tourism (Enter-2002), Springer, Innsbruck, Austria, (2002), 311-319.

[McEnery and Wilson 1996] McEnery, T. and Wilson, A.: "Corpus Linguistics", Edinburgh University Press, Edinburgh (1996).

[Perdue 1985] Perdue, R. R.: "Segmenting State Travel Information Inquiries by Timing of the Destination Decision and Previous Experience", Journal of Travel Research, 26, 4 (1985), 2-6.

[Roussinov and Zhao 2004] Roussinov, D. and Zhao, J. L.: "Automatic Discovery of Similarity Relationships through Web Mining", Decision Support Systems, 35, (2003), 149-166.

[Scharl 2004] Scharl, A.: "Web Coverage of Renewable Energy", Environmental Online Communication, A. Scharl (Ed.) Springer, London, (2004), 25-34.

[Scharl et al 2004] Scharl, A., Pollach, I. and Bauer, C.: "Determining the Semantic Orientation of Web-based Corpora", Intelligent Data Engineering and Automated Learning, 4th International Conference, IDEAL-2003, Hong Kong (Lecture Notes in Computer Science, Vol. 2690), J. Liu, Y. Cheung and H. Yin (Eds.) Springer, Berlin, (2003), 840-849.

[Sheldon 1993] Sheldon, P. J.: "Destination Information Systems", Annals of Tourism Research, 20, (1993), 633-649.

[Stone 1997] Stone, P. J.: "Thematic Text Analysis: New Agendas for Analyzing Text Content", Text Analysis for the Social Sciences, C. Roberts (Ed.) Lawrence Erlbaum, Mahwah, (1997), 35-54.