Information Technology & Tourism, Vol. 10 pp. 3–17
Printed in the USA.
1098-3058/08 $60.00 + .00

www.cognizantcommunication.com

ANALYZING NEWS MEDIA COVERAGE TO ACQUIRE AND STRUCTURE TOURISM KNOWLEDGE

ARNO SCHARL,* ASTRID DICKINGER,* and ALBERT WEICHSELBRAUN†

*Department of New Media Technology, MODUL University Vienna, Vienna, Austria
†Department of Information Systems and Operations, Vienna University of Economics & Business Administration, Vienna, Austria

Destination image significantly influences a tourist's decision-making process. The impact of news media coverage on destination image has attracted research attention and became particularly evident after catastrophic events such as the 2004 Indian Ocean earthquake that triggered a series of lethal tsunamis. Building upon previous research, this article analyzes the prevalence of tourism destinations among 162 international media sites. Term frequency captures the attention a destination receives—from a general and, after contextual filtering, from a tourism perspective. Calculating sentiment estimates positive and negative media influences on destination image at a given point in time. Identifying semantic associations with the names of countries and major cities, the results of co-occurrence analysis reveal the public profiles of destinations, and the impact of current events on media coverage. These results allow national tourism organizations to assess how their destination is covered by news media in general, and in a specific tourism context. To guide analysts and marketers in this assessment, an iterative analysis of semantic associations extracts tourism knowledge automatically, and represents this knowledge as ontological structures.

Key words: Knowledge acquisition; News media; Destination coverage; Sentiment analysis; Tourism ontology

Introduction

Trends in global tourism have shifted remarkably over the last decade. Information technology supports the increased sophistication of travelers (Chen & Sheldon, 1997), who seek greater variety in their travel arrangements and expect personalized services that meet their unique needs (Sheldon, 1993). Previously, travelers received destination information primarily through books, brochures, promotional videos, word-of-mouth, travel agents, or tourist offices. In recent years, marketers have observed a strong shift towards the World Wide Web as an additional and increasingly important source of destination information. This shift represents a challenge for destination managers, who have to promote a complex product in a highly competitive industry. To meet this challenge,

Address correspondence to Arno Scharl, MODUL University Vienna, Department of New Media Technology, Am Kahlenberg 1, 1190 Vienna, Austria. Tel: +43-1-3203555-500; Fax: +43-1-3203555-903; E-mail: scharl@modul.ac.at

many organizations focus on image studies, marketing strategies, conversion studies, and advertising research (Dore & Crouch, 2003), but often lack appropriate methods and tools to react quickly and effectively when confronted with unplanned and incidental news and media coverage.

In the past, the process of collecting media data was time consuming, expensive, and often resulted in outdated and incomplete information. Nowadays, detailed destination profiles and news media articles are readily available online, allowing for inexpensive, fast, and topical research. While still considered an emerging medium, the Internet has developed into an important source of destination information (Cai, Feng, & Breiter, 2004). Yet travelers often find it difficult to identify material of high quality, 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 (Cai et al., 2004; Guy, Curtis, & Crotts, 1990; Perdue, 1985). Therefore, destination marketers provide detailed descriptions and visual material to guide purchase decisions, assisting potential customers in determining the length of stay and the level of expenditure (Fesenmaier, 1994).

While destination marketers tailor and refine their own information offerings, they have limited control over the content of online news media. This is particularly evident in the case of negative coverage, which tends to fluctuate heavily due to political events, scientific discoveries, military operations, or natural disasters. A recent website analysis shows how one particular destination (Macau) is represented differently in various sources such as travel agents, official tourism websites, blogs, and online travel magazines (Choi, Lehto, & Morrison, 2007). The unpredictability and dynamic nature of Web resources call for automated approaches to capture and analyze their content. This article presents such an automated approach, investigating the prevalence and image of destinations in online news media as of April 2005. There are various definitions of *destination*, distinguishing the term from *origin* or *market*. The *Encyclopedia of Tourism* refers to a destination as, "a geographical unit visited by tourists . . . , a village or a town or a city, a region or an island or a country" (Cho, 2000, p. 144). This article uses the term destination in terms of geographic units (i.e., countries and cities that were included in the analysis).

The first step of analyzing destination coverage compares general and tourism specific references, as the image of a country as a tourism destination does not necessarily correlate with its overall profile. This comparison promises new insights into the structure of destination coverage, and allows investigating the relative importance of tourism for the economy of a particular country. The second step employs natural language processing techniques to build a network of semantic associations, and uses this network to extend and validate domain specific tourism ontologies.

This article contributes to existing literature in two ways. First, it shows how to analyze a destination's representation in online media—in terms of extent, sentiment, and associated topics. Analysts and marketers gain insights about different news media, and how their agenda impacts the decision of what and how to report. This type of information can guide and improve decision making, but lacks a structured representation for creating shared meaning. The article's second main contribution, therefore, is an automated method to create such structured representations. More specifically, the method extends and validates tourism ontologies based on the content of news articles or other large document collections.

Methodology

Tourism is among the leading applications of electronic commerce technology, with a high percentage of actors maintaining independent and increasingly sophisticated websites. General and domain specific search engines based on Web crawler technology are crucial for locating destination information on these websites. Among the main challenges with regard to achieving high search engine rankings are the often poor visibility of destination marketing organizations, inadequate content, limited control over third-party coverage, the favoring of well-known destinations, and unspecific search strategies of potential visitors. Further challenges for destinations are attitudes and

the motivational process of attitude changes reflected in public communication and the individual behavior (Kelman, 1958). Media and vehicle selection, reach and media scheduling are of equal importance (Rossiter & Danaher, 1998; Rossiter & Percy, 1987), but not the focus of this research.

To address the challenge of visibility, Delgado and Bowen (2004) call for destination portals that retrieve, match, and deliver information based on advanced Web crawlers and lexical-statistical processing of semantic information. The project presented in this article lays the foundation for such portals by aggregating fragmented tourism information (Maedche & Staab, 2002). The webLyzard crawling agent (www.weblyzard.com) mirrors a selection of international news media sites from the *Kidon.com*, *ABYZNewsLinks.com*, and *News Link.org* directories in weekly intervals. The sample comprises 162 sites from seven English-speaking countries: US (62), UK (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 breadth-first strategy to retrieve 50 megabytes of textual data from the highest hierarchical levels, as documents of lower hierarchical levels (e.g., historic newspaper archives) are rarely accessed by visitors (Scharl, Wöber, & Bauer, 2003).

Preserving the original site structure, the parsing component splits the retrieved textual data into sites, documents, and sentences. The hierarchical output file is encoded in the Extensible Markup Language (XML). The system then identifies and removes redundant copies of news headlines and noncontextual navigational elements such as menu items and news tickers, whose appearance on multiple pages would distort frequency counts—particularly in conjunction with contextual filtering as described below.

Economic Relevance of Destination Coverage

To measure the extent of news media coverage on particular destinations, a case-insensitive pattern-matching algorithm processed regular expression queries based on the names of countries, capitals (www.citypopulation.de), and major cities according to the TravelGIS (www.travelgis.com) database. Regular expressions are formalisms that describe sets of character strings without enumerating their elements (Friedl, 2002). In the case of *Austria*, for example, the system queried the tokenized output for the following terms: AUSTRIA, GRAZ, INNSBRUCK, LINZ, SALZBURG, and VIENNA.

Media Attention and Destination Image

The lack of local context limits the explanatory power of word frequency data (Biber, Conrad, & Reppen, 1998; McEnery & Wilson, 1996). Only measuring the number of occurrences neglects author attitude, for example, 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 taken from a tagged dictionary (Scharl, 2004; Scharl, Pollach, & Bauer, 2003). Such a dictionary contains a list of terms that are assigned additional (usually linguistic) attributes. In the context of this research, the tagged dictionary contained 4,400 positive and negative sentiment words from Harvard's General Inquirer (Stone, 1997). Reverse lemmatization increased the size of the dictionary by adding about 3,000 syntactical variations such as conjugations and gerund forms (e.g., complain\ra\complains, complaining, complained).

Table 1 shows that the most frequently mentioned country was the US, followed by Canada, Australia, the UK, and Iraq. This ranking reflects the Anglo-American news media sample and the geopolitical events of 2005. Countries rarely mentioned were Mayotte, Saint Pierre and Miquelon, Saint Lucia, Svalbard, Guadeloupe, and the Northern Mariana Islands. In terms of positive sentiment, Palau takes the lead, followed by San Marino, Bahrain, Saint Lucia, and Niue. Negative coverage concentrated on Iraq, Vietnam, Angola, Somalia, and Equatorial Guinea.

Co-occurrence analysis (Roussinov & Zhao, 2003) sheds further light on the rankings in Table 1, assuming that semantically related terms regularly appear in the same text segments. Co-occurrence analysis allows identifying the most important topics associated with the highest and lowest

Table 1
Country Rankings by Frequency and Sentiment

	Frequency	Sentiment
Frequent countries		
USA	144,079	0.121
Canada	63,548	0.112
Australia	43,052	0.133
UK	40,636	0.114
Iraq	35,231	–0.197
Ireland	23,869	0.091
France	19,416	0.110
China	18,167	0.093
New Zealand	11,970	0.114
Italy	11,002	0.094
Rare countries		
Faroe Islands	44	0.176
Pitcarin	40	0.016
Tokelau	30	0.075
Martinique	30	–0.070
N Mariana Islands	22	0.161
Guadeloupe	22	–0.055
Svalbard	20	0.072
Saint Lucia	19	0.270
Saint Pierre & Miquelon	15	0.123
Mayotte	13	0.164
Positive coverage		
Palau	181	0.325
San Marino	93	0.282
Bahrain	556	0.276
Saint Lucia	19	0.270
Niue	75	0.252
Saint Kitts & Nevis	96	0.245
New Caledonia	73	0.245
Micronesia	58	0.229
Gabon	87	0.195
Belize	218	0.195
Negative coverage		
Eritrea	177	–0.105
Kuwait	753	–0.107
Congo	1,542	–0.110
Kyrgyzstan	768	–0.123
Sudan	2,284	–0.143
Equatorial Guinea	201	–0.144
Somalia	566	–0.149
Angola	843	–0.156
Vietnam	3,241	–0.176
Iraq	35,231	–0.197

ranking countries in terms of sentiment. Limiting the consideration set of co-occurring terms to nouns by means of part-of-speech tagging (Abney, 1996) reduced memory consumption and improved both the throughput and the quality of results (part-of-speech tagging analyzes and annotates textual corpora to distinguish nouns from articles, verbs, adjectives, etc.).

Table 2 illustrates that only a subset 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), and past or current military conflicts (Vietnam, Iraq). While general keywords not related to tourism can be relevant for tourism research when describing processes that might impact destination image, analysts also require specific tourism data. To accommodate this requirement, the following section introduces a contextual filtering component that yields domain specific results.

Contextual Filtering

The results presented in the preceding section do not consider whether news articles mention destinations in a tourism context, which limits their interpretability. The problem can be addressed by contextual filtering, which confines the computation of term frequency to occurrences in documents that contain tourism-relevant terms. To identify tourism coverage, a pattern matching algorithm checked for the presence of at least one of the following terms (question marks instruct the pattern matching algorithm to treat the preceding character optionally):

```
accommodations?, backpack(ing|ers?),
bed (and|&) breakfasts?, camping, desti-
  nation
(images?|information|marketing|
  positioning),
holiday(s|ing|makers?)?, honeymoon
  (er)?s?, hos?tels?, hospitality, mo-
  tels?, national parks?,
nature( |-)?(trails?|parks?), sight
  ( |-)?see(ing|rs?)?, souvenirs?, tour
  (guides?| operators?), touris(m|ts?),
  travel(s|l?ing |l?ers?)?, vaca-
  tion(s|ing)?.
```

In a tourism context, the news articles analyzed for this study most frequently refer to the US, Canada, Australia, Ireland, and the UK. As mentioned above, this ranking reflects the composition of the Anglo-American news media sample.

The distinction between a general corpus and a

Table 2
Keywords for Countries Receiving Strong Positive or Negative Coverage

Country	Sentiment	Top 10 Keywords (Significance)
Palau	0.325	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	0.282	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	0.276	schumacher (63484), alonso (62221), ferrari (46198), parliamentary election (27730), renault (20414), barrichello (20394), trulli (18128), presidential election (16989), israeli (15460), prix (14891)
Angola	–0.156	marburg virus (27089), luanda (24510), uige (20396), mishawaka (15738), tri-central (15379), decatur (15229), congo (15203), elkhart (15144), vincennes (14372), ebola-like (14139)
Vietnam	–0.176	iraq (19858), war (17465), uq wire (14421), kerry (9720), hanoi (6069), shields (6068), vietnam war (5391), discusses (5099), zaoui (4884), bird flu (4666)
Iraq	–0.197	war (62321), us (37146), baghdad (30703), saddam (26917), troops (23437), blair (14994), weapons (14640), hussein (13588), soldiers (13582), occupation (10736)

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 helps disambiguate the meaning of a term, and thus increases the validity of results. The term CASABLANCA, for example, can refer to Morocco's capital as well as the famous movie starring Humphrey Bogart and Ingrid Bergman. This is usually referred to as *lexical ambiguity*. Analyses based on term frequencies cannot grasp the context required to determine the correct sense of the word. While it does not completely eliminate the problem, contextual filtering decreases the likelihood of encountering lexical ambiguity (referring to the above example, a tourism article is more likely to contain a city guide than a movie review).

Figure 1 compares media coverage by country, distinguishing between the sentiment of overall media coverage (Sent) and the sentiment expressed in specific tourism coverage (Sent-T). Light gray indicates the most positive coverage, while darker shading represents negative media opinion. This reveals contextual differences in sentiment. The overall coverage on Iran and Bolivia, for instance, is more negative than specific tourism coverage of these countries.

Proportion of Tourism Coverage by Destination

Of particular interest is the proportion of tourism coverage (Freq-T) compared to total media coverage (Freq). This ratio hints at the relative importance of tourism for a destination's economy. Two economic indicators published by the World Tourism Organization—tourism arrivals per capita (ApC) and international tourism receipts in US$ per capita (RpC)—allow validating this assumption (Table 3).

The Polynesian Islands including Niue, the Cook Islands, and French Polynesia, for example, are among the world's prime tourism spots. Thus, it is not surprising 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 general information include elections in Djibouti, Serbia and Montenegro's negotiations with the European Union, the visit of the Taiwanese president to Kiribati, the volatile political situ-

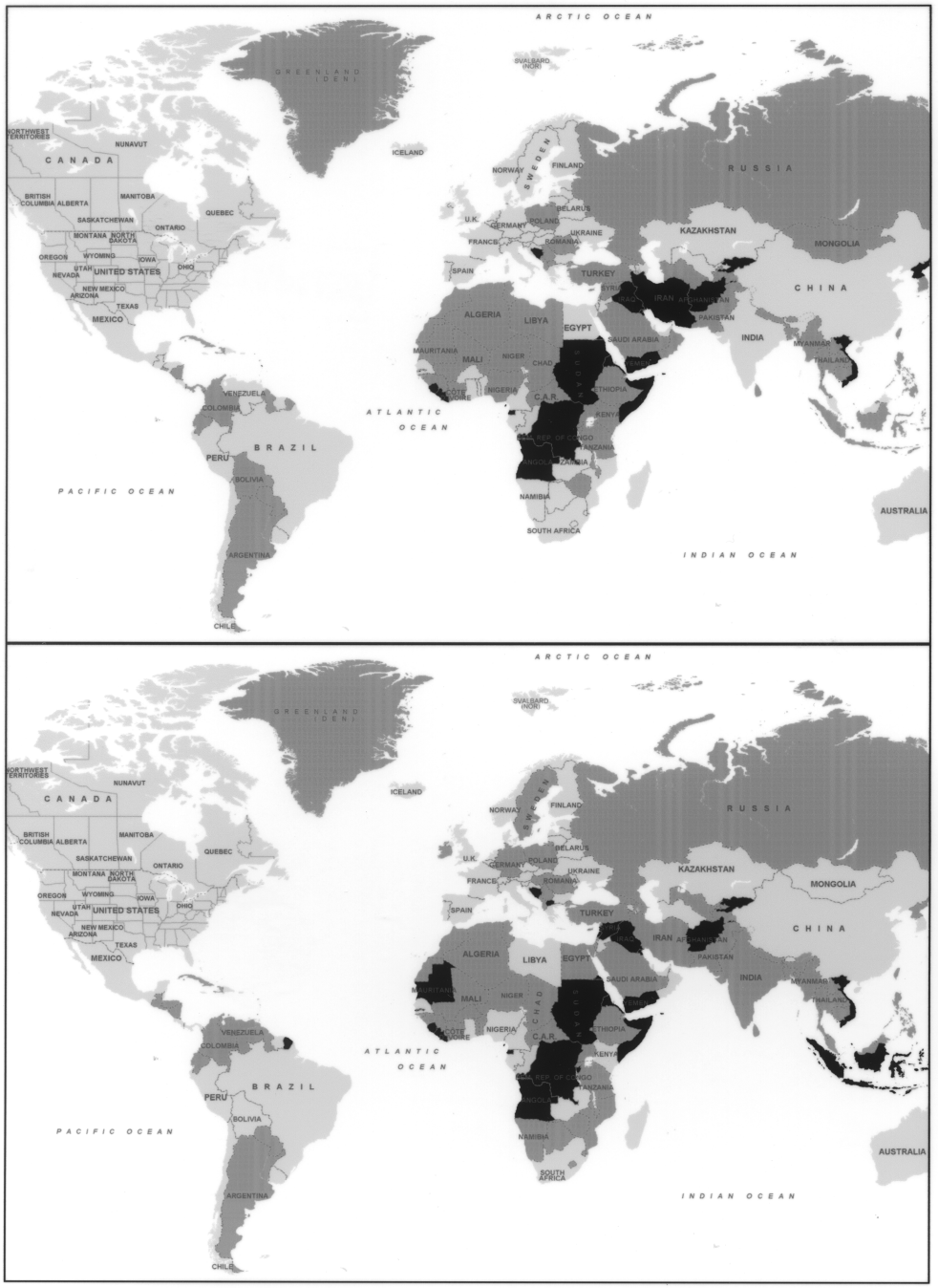

Figure 1. General versus tourism-specific news media coverage (top: general sentiment; bottom: sentiment in a tourism context); dark colors indicate negative coverage.

ation and foreign military use of air bases in Kyrgyzstan, and plans for a Russian space center at the Kourou Cosmodrome in French Guiana.

To test for correlations between sentiment, the percentage of tourism coverage, and the two economic indicators (ApC, RpC), the Spearman coefficient was used because not all variables showed normal distribution according to a Kolmogorov-Smirnov test. Highly significant at the 0.01 level, a correlation of 0.510** (0.412**) between the sentiment values and the arrivals (revenues) per capita supports the previously made proposition that there is a stronger relation between media *opinion* and travel decisions than the extent of coverage and travel decisions (the correlations between ApC/RpC and the general and tourism specific *frequencies* are low, and only one out of the four is significant at the 0.05 level; Freq-T/RpC with 0.156*).

While the amount of coverage has little predictive power, the proportion of tourism coverage relative to total coverage shows a positive, highly significant correlation with both arrivals per capita (0.421**) and tourism receipts per capita (0.390**). In other words, high proportions of tourism media coverage hint at the importance of tourism for a destination's economy, reflected in relatively higher arrivals and receipts per capita. Figure 2 illustrates two such regions with high proportion of tourism coverage, the regions of the Caribbean and Pacific Islands, both known as prime tourism destinations.

Planned marketing activities often initiate or at least influence positive tourism coverage on a country. 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 sentiment values for this group of islands, which normally is a favored destination. The low sentiment for the Caribbean island of Martinique represents an outlier due to reports

Table 3
Country Ranking by Tourism Coverage (in Percent of Total Coverage)

	Freq	Freq-T	Sent	Sent-T	Freq-%	ApC	RpC
MAX tourism coverage							
Niue	75	59	0.252	0.288	78.7	0.93	928
Dominica	93	72	0.168	0.171	77.4	1.05	736
Cook Islands	74	57	0.037	0.061	77.0	2.03	1,198
Maldives	303	221	0.066	0.065	72.9	1.66	937
Cayman Islands	191	138	0.129	0.104	72.3	6.82	13,572
Belize	218	154	0.195	0.153	70.6	0.81	487
N Mariana Islands	22	15	0.161	0.017	68.2	5.77	8,370
Martinique	30	20	–0.070	–0.051	66.7	1.04	570
French Polynesia	98	65	0.190	0.253	66.3	0.80	1,224
Netherlands Antilles	65	43	0.070	0.128	66.2	1.24	3,878
MIN tourism coverage							
Djibouti	82	10	0.030	–0.042	12.2	0.04	9
Kiribati	70	9	0.064	0.068	12.9	0.05	30
Serbia & Montenegro	1,835	244	–0.021	–0.058	13.3	0.04	7
Kyrgyzstan	768	119	–0.123	–0.205	15.5	0.01	5
French Guiana	212	34	–0.018	–0.061	16.0	0.34	235
Togo	305	52	0.080	–0.014	17.0	0.01	2
Moldova	209	36	0.160	0.340	17.2	0.00	12
Bahrain	556	96	0.276	0.203	17.3	0.04	929
Cóte d'Ivoire	48	9	0.053	–0.018	18.8	0.01	3
Suriname	68	13	0.089	0.113	19.1	0.13	32

Variables: Frequency and sentiment for the general (Freq, Sent) and the tourism corpus (Freq-T, Sent-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).

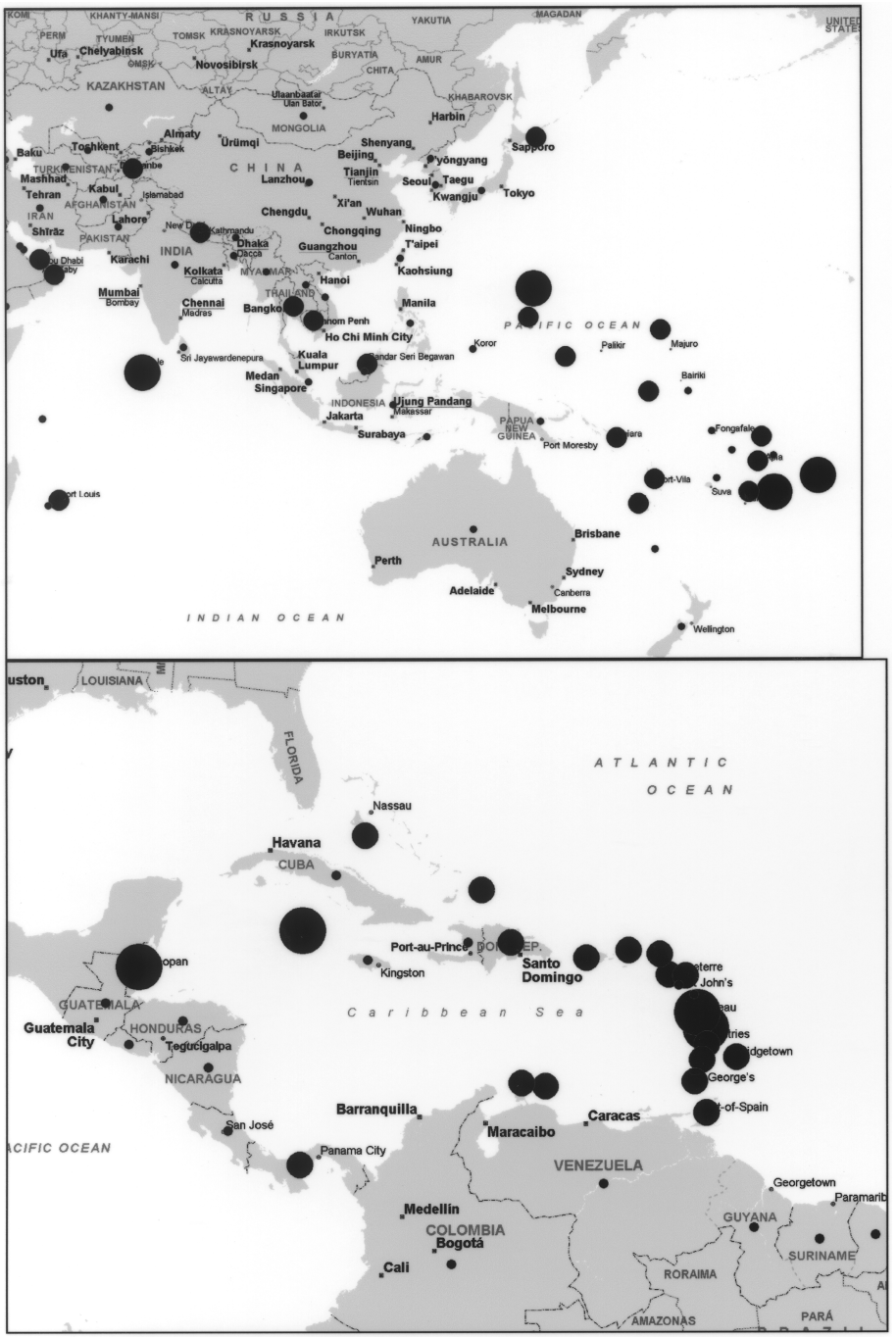

Figure 2. Tourism-related media coverage on the Pacific and Caribbean regions (the size of the black circular markers indicates the percentage of tourism coverage in relation to total media coverage).

on the 2004 crisis in neighboring Haiti and the anniversary of the 1961 death of Frantz Fanon, a West Indian psychoanalyst and social philosopher born in Martinique.

Automatically Extending and Validating Tourism Ontologies

Analyzing online media coverage contributes to effectively acquiring and managing tourism knowledge. But analysts and marketers who interpret results of automated content analysis, as exemplified in the preceding section, often rely on their implicit assumptions about the agenda (or the "world view") of different news media, and how this agenda might introduce a potential bias in terms of what they report on, and how they present current events to their audience.

The more sophisticated form of automated content analysis that is described in this section can replace these implicit assumptions by explicit formal representations based on empirical evidence. By iteratively analyzing semantic associations, tourism knowledge is extracted automatically, and represented as ontological structure. Thereby, it addresses one of the major research challenges of developing semantic services for the tourism industry (Dogac et al., 2004)—that is, the automated creation and validation of ontologies to establish a common understanding of concepts for humans and machines alike. While conflicting definitions of "ontology" abound (Guarino, 1998), there is consensus in the information systems literature that the term refers to a designed artifact formally representing shared conceptualizations (Gahleitner, Behrendt, Palkoska, & Weippl, 2005; Jarrar & Meersman, 2002). In the context of this research, ontologies are explicit formal specifications of terms used in the tourism domain, together with a set of hierarchical relations among them (Gruber, 1993). Specifying how these terms relate to each other, ontologies not only represent hierarchically organized knowledge, but also provide a common vocabulary for communicating about tourism related issues.

By providing shared meaning of words and concepts in specific knowledge areas (Fensel, Wahlster, Lieberman, & Hendler, 2003; Hjelm, 2001; Maedche & Staab, 2002), ontologies help integrate existing knowledge and support the next generation of data extraction, processing, and evaluation services.

Specifying a Seed Ontology

Several European research groups focus on developing tourism ontologies, acknowledging tourism as a data rich domain that draws upon numerous heterogeneous sources (Cardoso, 2006). The European *Harmonise* project and the follow-up *HarmoNET Tourism Harmonization Network* (www.harmo-ten.info), for example, provide data interoperability services that allow tourism organizations keeping their proprietary data formats while exchanging information via a mediator module based on a common tourism ontology (Dell'Erba, Fodor, Ricci, & Werthner, 2002). The Harmonise ontology currently comprises data items for accommodation, attractions and sights, events, and restaurants. Extending the Harmonise ontology by concepts relating to rural tourism, the *VMART* project deals with the availability and quality of information in the field of rural tourism micro enterprises (Richardson & Gudgeirsson, 2004).

Learning taxonomic relations from unstructured textual data (Cimiano, Pivk, Schmidt-Thieme, & Staab, 2005) is an important step in automating the creation and validation of such tourism ontologies. Initially, a small set of terms from domain experts or existing tourism ontologies is selected as seed ontology and formulated as a list of regular expressions. This research used the following seed ontology (the indentation reflects the ontology's hierarchical structure):

```
travel(s|l?ing|l?ers?)?
touris(m|ts?)
eco(-| )?touris(m|ts?)
cultur(?:al|e) touris(?:m|t|ts)
business travel(s|l?ing|l?ers?)?
commut(ers?|ing)
```

Hierarchical relations have been chosen as the primary focus of this research, acknowledging the importance of hierarchy in structuring human knowledge. As outlined in the concluding section, future research will automatically identify a range of different relation types.

Concept Identification and Positioning

The seed ontology terms are then fed into the *Lexical Analyzer*, which is the core of the ontology extension prototype. Figure 3 presents a conceptual view on the system architecture of this prototype (Liu, Weichselbraun, Scharl, & Chang, 2005).

Plurals, gerund forms, and past tense suffixes are syntactical variations that complicate the automatic processing of textual information. Lemmatizing the media corpus addresses this problem, putting verb forms into the infinitive, nouns into the singular, and removing elisions. This research used an adapted version of Someya's lemma list containing 40,569 words in 14,762 lemma groups (Someya, 1998). Lemmatizing the underlying corpus improves the ontology building process by grouping words of similar meaning, thereby increasing the stability and generalizability of the knowledge base.

Co-occurrence analysis at both the sentence and the document level then identifies semantically related terms (Roussinov & Zhao, 2003). Terms co-occurring on the sentence level tend to be more specific than those co-occurring on the document level. Candidate terms are selected according to a threshold value on the co-occurrence significance and checked against the WordNet lexical dictionary for word sense disambiguation (Navigli & Velardi, 2005). Further lexical analysis searches the Web corpus for terms connected by *trigger phrases* that indicate parent–child relations (Joho, Sanderson, & Beaulieu, 2004). In the phrase "specific tourism types *such as* ecotourism or cultural tourism," for example, SUCH AS indicates a hierarchical relation between the concept TOURISM TYPE as the superordinate parent, and ECOTOURISM and CULTURAL TOURISM as child nodes. Trigger phrases help determine which of two concepts is more general within a hierarchical structure.

The terms obtained are connected with the seed ontology via directed labeled links. The links' label includes term significance as determined by the co-occurrence analysis, and the method they originated from (co-occurrence on the sentence or document level, trigger phrases). Converting the semantic network by transforming labeled into weighted links through heuristic transformation rules yields the corresponding spreading activation network (a process that considers link type and significance level). By adjusting these transforma-

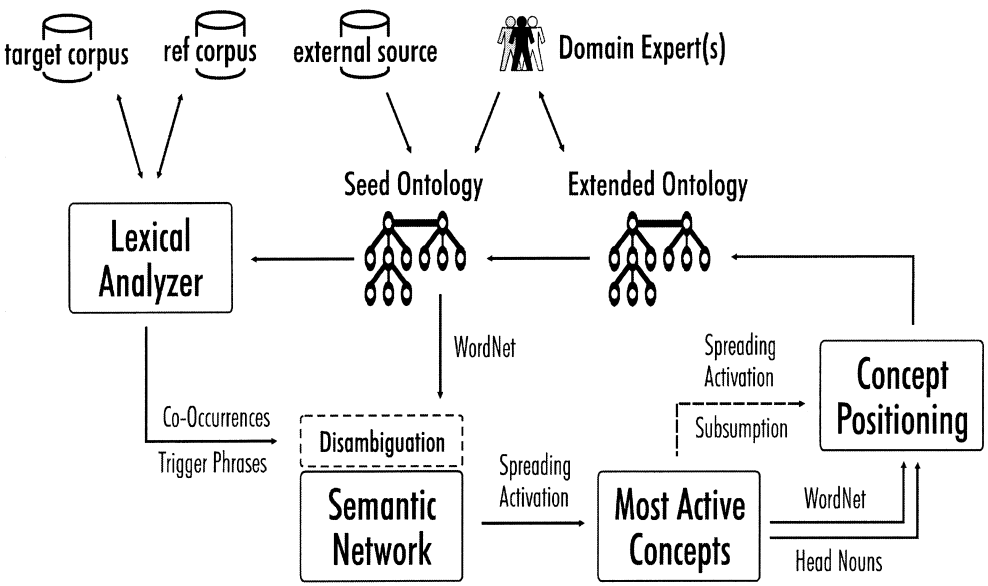

Figure 3. Ontology Extension System Architecture (Liu et al., 2005).

tion rules, domain experts can decide a priori whether more specific or general terms should be incorporated into the ontology.

Once the network is established, the system activates the seed concepts and identifies the most relevant domain terms via a spreading activation iteration. Only single common nouns or noun-noun combinations are considered for the candidate list, thus excluding proper nouns and other parts of speech. Grammatical relations, WordNet queries (Fellbaum, 1998), spreading activation, and subsumption analysis (Sanderson & Croft, 1999) then determine the semantic relation between concepts (the subsumption approach assumes that general terms occur more frequently than specific terms). Optionally, the system consults domain experts for terms not confirmed automatically, before the next iteration is triggered over the newly acquired terms.

The system outlined above facilitates the time-consuming process of eliciting and hierarchically positioning relevant domain concepts. It aims to accelerate the creation and diffusion of tourism ontologies, increase the completeness of the contained knowledge, and align ontology management with the requirements of tourism as a highly dynamic business environment (Pollach, Scharl, & Weichselbraun, 2007).

Extended Tourism Ontology

Figure 4 shows the extended tourism ontology after two iterations. Black nodes depict the seed ontology, while the gray and white nodes were added after the first and second iterations, respectively. Arrows indicate confirmed hierarchical relations. The broken lines connect semantically related terms whose exact type of relation could not be determined automatically. For these nonhierarchical relations, the (r) values indicate their strength based on the link assignment's spreading activation level. High values suggest a strong relation between the concepts, a value of 8 being the maximum due to the specific setup of the spreading activation network.

The seed ontology included six concepts: travel, tourism, commuter, business travel, eco-tourism, and culture tourism. Two iterations added 24 terms, yielding a representation of the most relevant concepts that the news media associated with the seed ontology's concepts. They were not only automatically identified but also grouped into hypo- and hypernyms whenever the identification of hierarchical structure was possible. Some of the hierarchical relations directly stem from the WordNet dictionary, which for example lists flight and trek as hyponyms of trip. It is interesting to note that all but one of the 12 terms added after the first iteration are linked to travel, as this seed term is the most generic and represents a fundamental part of tourism (Wall, 2000).

Concepts connected to the seed term TRAVEL belong to the fields of transportation (AIRLINE, AIR TRAVEL, FLIGHT), travel industry (TRAVEL BUSINESS, TRAVEL AGENCY), and travel destination (DESTINATION, HOTEL). Among the strongest relations are those between the seed term TRAVEL and three terms added in the first iteration: TRAVEL BUSINESS ($r = 4.9$), TRAVEL AGENCY ($r = 4.8$), and DESTINATION ($r = 4.8$). TRAVEL BUSINESS was associated with two further concepts in the second iteration (BUSINESS and RAIL). While the relations seem intuitive, the specific modes of transportation identified (RAIL and AIR TRAVEL) remain unconnected.

The airline industry is represented by the terms AIRLINE, AIR TRAVEL, FLIGHT, and AIRPORT. The latter was correctly added in the second iteration. The system connected all those terms to the seed ontology's TRAVEL, and established an additional direct link between FLIGHT and AIR TRAVEL, resulting in a triangular relation. As the airline industry is among the strongest players in electronic commerce, it does not surprise that online sources emphasize this particular sector. Only the association between SNOW CAM and AIR TRAVEL remains unclear in the first place, but further investigation of the raw data showed that the link was caused by coverage on different forms of travel including air travel, mentioning 10 snow cams in different locations.

The four terms AUDIO TOUR, ART FESTIVAL, NATION CULTURE, and HANDCART relate to CULTURE T\OURISM ($r = 8.0$), but not hierarchically. While the first three terms are plausible associations, the fourth term HANDCART represents an outlier that

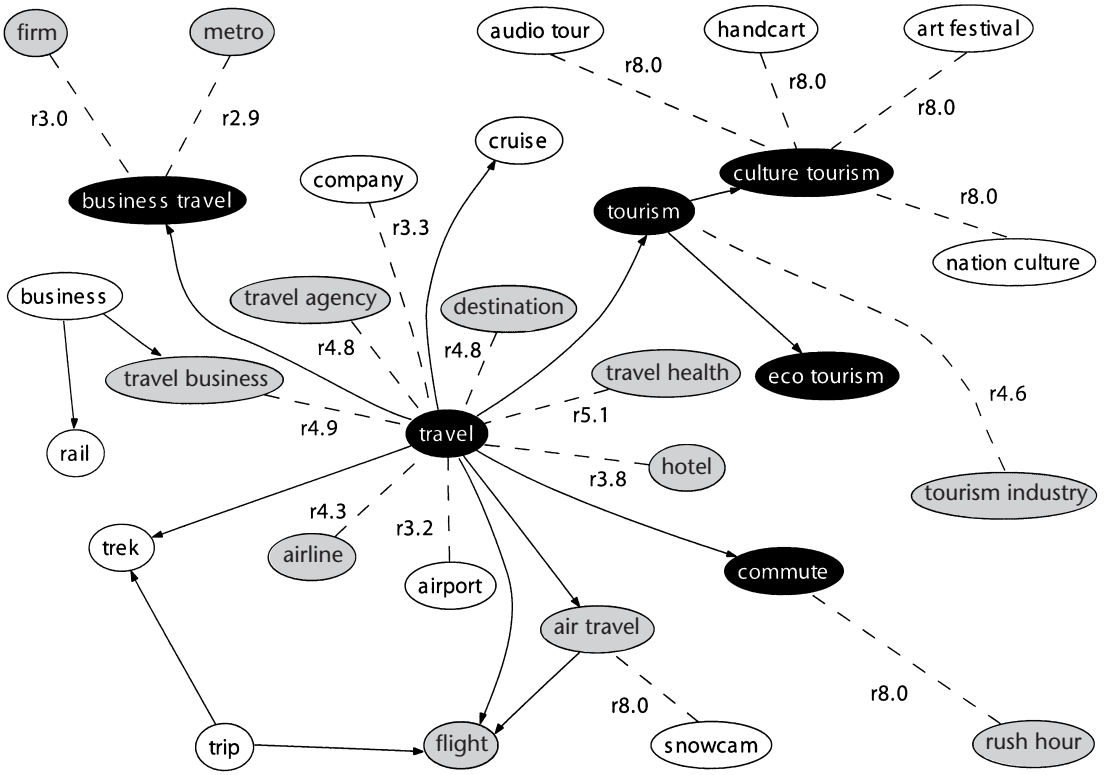

Figure 4. Hierarchy of tourism concepts after two iterations.

was added due to CNN and USA Today coverage on a "Mormon Handcart Track," referring to cultural tourism.

A strong relation was found between commute and rush hour ($r = 8.0$). Business travel also received two associated concepts: metro and firm. Alternatively, a human analyst might have connected METRO to COMMUTE and RUSH HOUR. For the algorithm, however, those links were less significant as METRO not only relates to the urban transportation system, but also forms part of company names (e.g., real estate agencies or convention centers). This strengthens the association between BUSINESS TRAVEL and METRO, which is misleading in this particular case. An improved proper noun detection that disregards elements of composite proper nouns will help avoid this problem. Similar difficulties arise from terms with multiple meanings—for instance FIRM in the sense of "determined" versus FIRM representing an organizational entity. Our current approach counters these difficulties by only considering nouns (eliminating non-noun senses) and strictly enforcing the seed ontology context (eliminating most out-of-context senses). These measures are computationally efficient and straightforward to implement, but do not eliminate interference by terms with multiple meanings completely. A refined algorithm should handle concepts as separate entities, and provide full word disambiguation on the concept level.

Conclusions and Future Research

Media coverage influences the image of tourism destinations. The automated analysis of tourism-related media coverage helps investigate this influence by revealing the public profile of particular destinations, as well as the impact of current events. Computing sentiment 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.).

Creating shared meaning is the main motivation for building tourism ontologies (Fensel et al., 2003; Maedche & Staab, 2002). The ontology extension prototype presented and applied in this article represents an innovative approach to building domain-specific ontologies. It helps distinguish between synonym–antonym and hyponym–hypernym pairs, extends and validates tourism knowledge, and sheds light onto the representation of tourism-related issues in news media articles. Automated ontology extension is typically applied in conjunction with human expertise, as its goal is not to provide a universally correct representation of a domain, but a snapshot of knowledge contained in a specific corpus of text that might comprise millions of documents and, thus, eludes manual analyses. Additional sources of destination information (RSS/Atom news feeds, websites of national tourism organizations, travelers' blogs, etc.) will improve the quality of such as snapshot, and allow for a fine grained semantic disambiguation. Automatic knowledge processing requires nonambiguous terminology, but many tourism terms such as ecotourism or heritage tourism leave room for diverging and often conflicting interpretations.

From a methodological perspective, future research will investigate the automated discovery of relation types, as well as the evolution of ontologies based on dynamically changing corpora. This addresses the critical questions of how different types of organizations present a certain topic over time, how the covered time span affects the stability of ontological descriptions, and how the choice of seed terms impacts ontology evolution.

From an applied perspective, future research will integrate ontology knowledge to disambiguate and refine queries, build context-aware information retrieval agents, and improve the tracking and benchmarking of destinations in online environments. Just-in-time information retrieval agents (Rhodes & Maes, 2000), for example, can use ontology knowledge to identify relevant material automatically, and to proactively retrieve and present this material in an easily accessible yet nonintrusive manner. Finally, introducing a longitudinal perspective and tracking destination coverage over time will allow distinguishing between superficial changes in attitude on the verbal level from lasting changes firmly integrated into the authors' value systems (Kelman, 1958).

Acknowledgment

The IDIOM (Information Diffusion across Interactive Online Media; www.idiom.at) research project is funded by the Austrian Federal Ministry of Transport, Innovation & Technology (BMVIT) and the Austrian Research Promotion Agency (FFG) within the strategic objective FIT-IT Semantic Systems (www.fit-it.at).

Biographical Notes

Prof. Arno Scharl heads the Department of New Media Technology at MODUL University Vienna. Prior to his current position, he held professorships at Graz University of Technology and the University of Western Australia, as well as a visiting fellowship at the University of California at Berkeley. His current research focuses on integrating semantic and geospatial Web technology, computer-mediated collaboration, ontology learning, sustainability, and environmental communication.

Dr. Astrid Dickinger is Assistant Professor at the Department of New Media Technology at MODUL University Vienna. Previously she was Assistant Professor at the Institute for Tourism and Leisure Studies at Vienna University of Economics and Business Administration where she completed her dissertation. Her research interests are in the areas of service quality, models of consumer behavior, acceptance of innovations, electronic, and mobile service usage.

Dr. Albert Weichselbraun is Assistant Professor at the Vienna University of Economics and Business Administration's Department of Information Systems and Operations. In 2004 he joined the webLyzard project, which develops Web mining tools to gather, aggregate and analyze unstructured textual data from Web resources. His current research focuses on ontology evolution, automated ontology learning, and the analysis of semantic, social, and geospatial relations.

References

Abney, S. (1996). Tagging and partial parsing. In K. Church, S. Young, & G. Bloothooft (Eds.), *Corpus-based methods in language and speech* (pp. 118–136). Dordrecht: Kluwer Academic Publishers.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus lin-*

guistics—investigating language structure and use. Cambridge: Cambridge University Press.

Cai, L. A., Feng R., & Breiter, D. (2004). Tourist purchase decision involvement and information preferences. *Journal of Vacation Marketing, 10*(2), 138–148.

Cardoso, J. (2006). Developing an OWL ontology for e-tourism. In J. Cardoso & A. Sheth (Eds.), *Semantic web services, processes, and applications* (pp. 247–282). Berlin/Heidelberg: Springer.

Chen, H. M., & Sheldon, P. J. (1997). Destination information systems: Design issues and directions. *Journal of Management Information Systems, 14*(2), 151–176.

Cho, B. H. (2000). Destination. In J. Jafari (Ed.), *Encyclopedia of tourism* (pp. 144–145). London: Routledge.

Choi, S., Lehto, X. Y., & Morrison, A. M. (2007). Destination image representation on the web: Content analysis of Macau travel related websites. *Tourism Management, 28*, 118–129.

Cimiano, P., Pivk, A., Schmidt-Thieme, L., & Staab, S. (2005). Learning taxonomic relations from heterogeneous evidence. In. P. Buitelaar, P. Cimiano, & B. Magnini (Eds.) *Ontology learning from text: methods, applications, and evaluation*. Amsterdam: IOS Press.

Delgado, J. A., & Bowen, M. (2004). DestinationFinder: A travel focused search engine, portal, and recommender system for the DMO. In A. Frew (Ed.), *Information and communication technologies in tourism 2004* (Enter-2004), (CD Rom) Cairo, Egypt.

Dell'Erba, M., Fodor, O., Ricci, F., & Werthner, H. (2002). Harmonise: A solution for data interoperability. In *IFIP Conference Proceedings, Towards the Knowledge Society: E-Commerce, E-Business, E-Government* (Vol. 233). Lisbon, Portugal: International Federation for Information Processing.

Dogac, A., Kabak, Y., Laleci, G., Sinir, S., Yildiz, A., Kirbas, S., & Gurcan, Y. (2004). Semantically enriched web services for the travel industry. *SIGMOD Record, 33*(3), 21–27.

Dore, L., & Crouch, G. I. (2003). Promoting destinations: An exploratory study of publicity programmes used by National Tourism Organisations. *Journal of Vacation Marketing, 9*(2), 137–151.

Fellbaum, C. (1998). WordNet an electronic lexical database. *Computational Linguistics, 25*(2), 292–296.

Fensel, D., Wahlster, W., Lieberman, H., & Hendler, J. (2003). *Spinning the semantic web—bringing the World Wide Web to its full potential*. Cambridge: MIT Press.

Fesenmaier, D. R. (1994). Traveller use of visitor information centers: Implication for development in Illinois. *Journal of Travel Research, 33*(1), 44–50.

Friedl, J. E. F. (2002). *Mastering regular expressions*. Sebastopol: O'Reilly Media.

Gahleitner, E., Behrendt, W., Palkoska, J., & Weippl, E. (2005). Knowledge sharing and reuse: On cooperatively creating dynamic ontologies. In *16th ACM Conference on Hypertext and Hypermedia*, Salzburg, Austria.

Gruber, T. R. (1993). A translation approach to portable ontology specification. *Knowledge Acquisition, 5*(2), 199–220.

Guarino, N. (1998). Formal ontology and information systems. In N. Guarino (Eds.), *First international conference on formal ontologies in information systems (FOIS-98)* (pp. 3–15). Trento, Italy: IOS Press.

Guy, B. S., Curtis, W. W., & Crotts, J. C. (1990). Environmental learning of first-time travelers. *Annals of Tourism Research, 17*(3), 419–431.

Hjelm, J. (2001). *Creating the semantic web with RDF: Professional developer's guide*. New York: Wiley.

Jarrar, M., & Meersman, R. (2002). Formal ontology engineering in the DOGMA approach. In R. Meersman & Z. Tari (Eds.), *International Conference on Ontologies, Databases, and Applications of Semantics* (Lecture Notes in Computer Science, Vol. 2519) Berlin: Springer.

Joho, H., Sanderson, M., & Beaulieu, M. (2004). A study of user interaction with a concept-based interactive query expansion support tool. In *Advances in Information Retrieval, 26th European Conference on Information Retrieval*.

Kelman, H. C. (1958). Compliance, identification, and internalization three processes of attitude change. *The Journal of Conflict Resolution, 2*(1), 51–60.

Liu, W., Weichselbraun, A., Scharl, A., & Chang, E. (2005). Semi-automatic ontology extension using spreading activation. *Journal of Universal Knowledge Management, 0*(1), 50–58.

Maedche, A., & Staab, S. (2002). Applying semantic web technologies for tourism information systems. In *Information and Communication Technologies in Tourism 2002 (Enter-2002)* (pp. 311–319). Wien/New York: Springer.

McEnery, T., & Wilson, A. (1996). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Navigli, R., & Velardi, P. (2005). Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(7), 1075–1086.

Perdue, R. R. (1985). Segmenting state travel information inquiries by timing of the destination decision and previous experience. *Journal of Travel Research, 26*(4), 2–6.

Pollach, I., Scharl, A., & Weichselbraun, A. (2007). Web content mining for comparing corporate and third-party online reporting: A case study on solid waste management. *Business Strategy and the Environment, 15*.

Rhodes, B. J., & Maes, P. (2000). Just-in-time information retrieval agents. *IBM Systems Journal, 39*(3/4), 685–702.

Richardson, P., & Gudgeirsson, G. (2004). Solving problems in rural tourism using semantic web technologies. In A. Frew (Ed.), *Information and communication technologies in tourism 2004* (Enter-2004) (CD Rom), Cairo, Egypt.

Rossiter, J. R., & Danaher, P., J. (1998). *Advanced media planning*. Boston/Dordrecht/London: Kluwer Academic Publishers.

Rossiter, J. R., & Percy, L. (1987). *Advertising & promotion management*. New York: McGraw-Hill Book Company.

Roussinov, D., & Zhao, J. L. (2003). Automatic discovery

of similarity relationships through web mining. *Decision Support Systems, 35*, 149–166.

Sanderson, M., & B. W. Croft (1999). *Deriving concept hierarchies from text*. Paper read at 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, USA.

Scharl, A. (2004). Web coverage of renewable energy. In A. Scharl (Ed.), *Environmental online communication* (25–34). London: Springer.

Scharl, A., Pollach, I., & Bauer, C. (2003). Determining the semantic orientation of web-based corpora. In J. Liu, Y. Cheung, & H. Yin (Eds.), *Intelligent Data Engineering and Automated Learning, 4th International Conference, IDEAL-2003*, Hong Kong (Lecture Notes in Computer Science, Vol. 2690) (pp. 840–849). Berlin: Springer.

Scharl, A., Wöber, K. W., & Bauer, C. (2003). An integrated approach to measure web site effectiveness in the European hotel industry. *Information Technology & Tourism, 6*(4), 257–271.

Sheldon, P. J. (1993). Destination information systems. *Annals of Tourism Research, 20*, 633–649.

Someya, Y. (1998). *e_lemma.txt*. Retrieved January 1, 1999, from http://www.lexically.net/downloads/e_lemma.zip

Stone, P. J. (1997). Thematic text analysis: New agendas for analyzing text content. In C. Roberts (Ed.), *Text analysis for the social sciences* (pp. 35–54). Mahwah: Lawrence Erlbaum.

Wall, G. (2000). Travel. In J. Jafari. *The encyclopedia of tourism* (pp. 600/n/601). London: Routledge.