JOURNAL OF H&T
www.johat.org

Knowledge Transfer in the Tourism Industry: Implications of the Use of Social Media in Business Analysis

Malcolm Cooper
Le Quang Thai
William Claster
Kazem Vafadari
Phillip Pardo
Ritsumeikan Asia Pacific University, Japan

Abstract
Attracting tourists to a destination through favorable imagery is an essential part of the transfer of knowledge in the tourism and hospitality industry. Destination image is defined as the expression of objective knowledge, imagination and the subjective emotional thoughts of an individual or group. In turn, social media is profoundly changing not only the way we communicate, but also the way human beings interact with this imagery and their travel environment. Therefore, in order to become more competitive, many firms are seeking to leverage the power of social media to gain insights into customer cognition and behavior. Social media is beginning to be used for in predicting customer behavior in a wide range of industries, including restaurants, retailing, and the hotel industry. By analyzing a collection of search engine queries, and historical Twitter data, this paper investigates the competitive intelligence of two different groups of hotels in Philadelphia, USA. Further the paper provides an analysis of the correlation between social media and hotel occupancy rates, and proposes a methodology to predict future lodging demand from empirical data in line with the objectives of the t-Forum. Our previous research has shown that mining of social media can be used to explain and forecast social and business trends in tourism. In this paper, we extend this research to evaluate the transfer of knowledge in the industry. It is hoped that our comments will enable t-forum to come up with advice on the transfer of knowledge about destinations.
Key words: Social media, Knowledge transfer, Search engines, Hotel industry, Competitive intelligence, Twitter

Introduction
Social networking media such as Facebook, Twitter, and Google Plus are becoming more prevalent and, as a result human beings are changing the way they receive and share information. As a result, by adopting social media networking, businesses and organizations worldwide are now able to communicate more easily with their customers in both directions. Sending strategic messages to the market, and at the same time retrieving feedback from customers in real time, is profoundly changing our way of doing business. In fact, social media can not only enhance knowledge regarding customers, but it can also offer insight into competing organizations. Specifically, companies can now gain competitive intelligence by

discovering interesting knowledge and patterns from textual online data; in order to understand more fully products, customers, and competitors, and then converting these data into intelligence that can support business decision-making (Dey, Haque, Khurdiya & Shroff, 2011).

With respect to the hotel industry however, despite the fact that there have been some efforts in adopting social media for guest analysis purposes. Starwood Hotels and Resorts was one of the first hotel groups to utilize social media to support potential guests in making travel decisions (Lanz, Fishhof & Lee, 2010), however there is still a lack of comprehensive analysis on how social media affects the lodging market (Anderson, 2012). In an effort to support the lodging industry, and perhaps assist the t-Forum bid to enhance knowledge transfer in the industry, this paper focuses on the use of various data mining techniques to use Google search engine queries and the sentiment mining of tweet data to analyze the competitive intelligence of two groups of hotels in Philadelphia, USA. We concentrate on 5-star brands as they are more easily identifiable in the data stream, but in time we hope that the methodologies outlined in the paper will be used to gain insights into the relationship between social media data and hotel occupancy rates across the board (Blal & Sturman, 2014).

The objective of the present research was simply to examine whether tweets can be sentiment mined for market intelligence on tourism facilities and attractions; in other words, the views being expressed by actual and/or potential consumers in a reliable and as close to real-time way. The results lead us to propose that micro-blogs can be used as a source of useful information in the development of user-provider relationships within the accommodation industry. We came to this conclusion through capturing and studying the sentiments expressed in micro-blogs related to a range of different tourism facilities destinations (reporting on two of these in this paper), investigating the ways sentiment mining could be used in market research for the tourism industry. In achieving its objective, this paper reviews the current research undertaken related to sentiment mining and tourism, explains the technological background, the nature and location of the data and the analytical models employed. We also discuss the reliability of these models. Finally conclusions and directions for future work are described.

Background

In 2007 the term 'social technographics' was coined by the Forrester Research Group in the US (since shortened to sociographics) in order to help businesses engage in social media with a more human approach, catering to individuals where, when, and how they are participating and contributing to the social Web (Forrester Group, 2007). By 2010 the group was reporting that one in every three online Americans is a "conversationalist" - someone who updates their status in the statusphere (any social network with an update window) at least once per week. Conversationalists represent 33% of today's online social behavior. The goal of sociographics was however not only to classify individual participation in social media, but also to encourage the design and segmentation of focused marketing, branding, and engagement programs that would appeal to these respective groups. The Group believed that in order to effectively form ties that bind with customers, businesses must genuinely understand the social behaviors of consumers. By personalizing the messages and the digital conduits between brands and markets, businesses evolve from a carpet bombing campaign that is essentially marketed at faceless consumers toward using mediums that appeal to targeted demographics (characterized by age, income, gender, education, etc.), instead of psychographics (grouped by interest).

The tourism industry in any location is highly dependent on customer sentiment and tourist perceptions of currently available services for its capture of visitor flows. However, the cost of monitoring these perceptions in near real-time for appropriate action can be prohibitively expensive, and may often not be possible with traditional methods of analysis of tourism motivations and targeted demographics. Since the introduction of the Internet , enormous opportunities have opened up for social interaction at a distance (Forrester Group, 2007; Akehurst, 2009). Tourism in general is ranked one of the leading industries in terms of online transactions (Werthner & Ricci, 2004), and is well placed to take advantage of these new opportunities revealed by sociographics (Adam et al., 2007). However *actual* on-line content may be limited in value and relatively difficult or costly to locate (Akehurst, 2009). It may in fact involve complex ongoing relationships revolving around lifestyle issues and the value placed on such sources of information. The barriers to effective Internet usage according to Carson (2005) include technical competence, variations in technology, adoption by governments, enterprises and consumers, resistance to the innovations brought on by the Internet, access to IT infrastructure, the costs of using IT, and the existence of government policies which might support but equally might discourage effective Internet exploitation.

So while access to hard-to-reach market segments, or uncovering the unsuspected strengths and weaknesses of a particular tourism destination or organization may now be possible using the internet, the challenge for market researchers and managers is how to search and visit the vast number of social media outlets in order to derive up-to-date and useful information (Akehurst, 2009). Choi et al. (2007) relied on analyzing the material contained within online magazines and full length blogs, while in our work, we take advantage of data that is both delivered in real-time and time-date stamped through *micro-blogs* (tweets). Carson (2008) highlight the considerable time and effort required to find relevant information in blogs as well as the opportunities and usefulness of these sources.

There are many types of social media data, but most of them are in forms of contextual and unstructured text data such as posts, comments on Facebook, Twitter, hotel reviews on Trip Advisor, or search query index from Google, Baidu, and so on (Hepburn, 2007). Over the last few years, social media data analysis has become a trend in business intelligence. To illustrate, Google search queries relating to travel was used to forecast the effects of oil spills on lodging demand in the Persian Gulf region (Choi & Liu, 2011), or predict tourist volumes in Hainan Province, China (Yang, Pan, Evens & Lv, 2014). Moreover, text-mining techniques have also been used to facilitate the exploitation of patterns of competitive intelligence from Twitter data in the Pizza industry (He, Zha & Li, 2013), as well as sentimental knowledge about the development of customer relationships within the hospitality industry (Claster, Pardo, Cooper & Tajeddini, 2013). Additionally, there has been a rapid development in the tools and algorithms enhancing social media data analysis; such as SAS text analysis, Alchemy API, Rapid Miner, and so on.

Other studies show that there has been a considerable amount of work recently undertaken in sentiment mining (Pang & Lee, 2008; Barbier & Liu, 2011). The results show that the limited size of twitter micro-blogs (140 words) may actually boost text mining efficiency, as subject ambiguity is reduced in these shorter expressions, although conclusive research has not been conducted (Jindal and Liu, 2006). Jansen, Zhang, Sobel & Chowdury (2009), for example, analyzed 150,000 micro-blogs from twitter in terms of frequency, timing, and contents of tweets within a corporate account and focused on the sentiment expressed towards products produced by that company. Go, Huang & Bhayan (2009) built an algorithm to classify sentiment within tweets as positive or negative and achieved accuracy as high as 81%, Kennedy and Inkpen (2006) examined the effect of 'valence shifters' such as negations, intensifiers, and 'diminishers' in modifying sentiment and extended the study of

UGCs by looking beyond unigram features to bigrams. Zhuang et al. (2006) mined the Internet movie database (IMDB) to derive sentiment scores and reported that the sentiment classification methodology compares favorably to an earlier methodology developed by Hu and Liu (2004) for summarizing customer reviews.

In the accommodation industry, the impact of blogs and in particular interactive and reader comments has increased along with the popularity of exchange, recommendation and opinion sites. Over 20% in 2006, and probably more today, of consumers (tourists) relied on user generated content (especially word of mouth) or UGC (Sarks, 2007), probably because of their need to obtain comments and opinions from those who have used a hotel before (Rabanser & Ricci, 2005; Senecal & Nantel, 2004), and the fact that they may believe in social media comments more than in travel agencies (Emarketer, 2007). It is worth noting that looking at other tourists' comments and travel blogs are popular online activities in relation to accommodation). A comprehensive study by ComScore (2007) confirmed that user reviews had a significant influence on customer purchases, and that reviews generated by fellow customers have a greater impact than those generated by professionals. In addition, online reviews are fast becoming an increasingly useful and important information resource as noted by Horrigan (2008), in that 81% of Internet users have done online research on a product at least once. Micro-blogging, Twitter being the best-known currently although other sites exist, takes the concepts of blogs to a much higher level of intimacy and immediacy between individuals, which encourages participation (Akehurst 2009; Beaumont 2008) and involvement.

Assessing and measuring the sentiment accumulated in the vast store of blogs, online publications, social network media and micro-blogs can thus yield tangible and actionable information for business, marketing, social sciences, and government. Previous research shows that through the use of blog content businesses can improve customer profiling, customer acquisition, customer engagement, brand awareness, brand reinforcement, reputation management and customer service (Dellarocas, 2003; Laboy & Torchio, 2007). Better knowledge of consumer opinions, public attitudes, and generally the “wisdom of crowds” can yield highly valuable information for business (Claster et. al., 2010). Moreover, as the World Wide Web has developed, considerable decision-making power over the use of discretionary products like accommodation services has been transferred from suppliers to consumers. There is a real need to improve market intelligence and market research for hotel organizations in order to facilitate timely decision-making and market orientation (i.e. customer orientation, competition orientation and inter-functional coordination), and in general, performance in the accommodation industry (Blal and Sturman, 2014). This requirement can be said to be universal in the entire tourism industry.

Study Methods and Analysis

Data Collection

This research makes use of data from various sources including Google trends data, Philadelphia's hotel industry related tweets data, and the average occupancy rate of 8 hotels in 2 groups in Philadelphia (Table 1).

Table 1. Hotels in the Study

Group	Hotel Name
5-Star International	Hyatt at the Bellevue
	Windsor Suites
	Le Meridien Philadelphia
	Kimpton Hotel Palomar Philadelphia
5-Star	Sofitel Philadelphia
	Four Seasons Hotel Philadelphia
	The Latham Hotel
	AKA Rittenhouse Hotel

Google Trends Data: Google Trends is Google's online service and provides an index of the volume of Google searches that can be classified into different locations (worldwide level or specific country), and topic categories (Travel, Society, Business, etc.). After providing a certain query string, the website outputs the normalized query volume over a timed period (daily, weekly, monthly or annually); filtered by selected location and category options. Google trends data on the selected hotels in Philadelphia from the year 2012 to 2014 was collected using the respective input queries as shown in Table 2. All of the collected information was filtered using "Worldwide" location and "Travel" categories.

Google Trends is Google's online service and provides an index of the volume of Google searches that can be classified into different locations (worldwide level or specific country), and topic categories (Travel, Society, Business, etc.). After providing a certain query string, the website outputs the normalized query volume over a timed period (daily, weekly, monthly or annually); filtered by selected location and category options. Google trends data on the selected hotels in Philadelphia from the year 2012 to 2014 was collected using the respective input queries as shown in Table 2. All of the collected information was filtered using "Worldwide" location and "Travel" categories.

Tweet postings on each of the eight selected hotels Twitter accounts:
According to Twitter policy, the latest streaming tweets data of a single Twitter user can be retrieved by the provided Application Programming Interface (API), at a maximum of 3200 tweets, and with various other limitations such as on the number of requests per 15-minute period. To do this we developed a web scraping program in the Ruby programming language to move tweet data to a local computer from an external online service called Tweet Tunnel, which shows up to 3200 of the latest tweets directly on its website. The tweets data consists of the Tweet ID, the username, tweet contents, and the posted date. The column "Number of Retrieved User's Tweets" in Table 2 illustrates the number of successful retrieved tweets for each of the eight selected hotels.

Table 2. Input Query Strings used to collect Google Trend data and the Tweet Data Collecting Result

Hotel Names	Input Query String on Google Trend	Number of Retrieved User Tweets
Hyatt at the Bellevue	Hyatt at The Bellevue	352/353 (99.7%)
Windsor Suites	Windsor Suites Philadelphia	3035/3037 (99.9%)
Le Meridien Philadelphia	Le meridien Philadelphia	694/694 (100%)
Kimpton Hotel Palomar Philadelphia	hotel Palomar Philadelphia+Palomar Philadelphia	2006/2011 (99.7%)
Sofitel Philadelphia	Sofitel Philadelphia	3151/6868 (46%)
Four Seasons Hotel Philadelphia	Four Seasons Philadelphia	3149/9081 (35%)
The Latham Hotel	The Latham Hotel + The Latham Philadelphia	40/40 (100%)
AKA Rittenhouse Hotel	Aka Rittenhouse Philadelphia	2924/2929 (99.8%)

Public Tweets: several aggressive searches were carried out on a personal local database of roughly 1 Tera bytes of public tweets. This was done to filter-out tweets talking about a certain member of the 8 selected hotels, or about the general lodging interest in Philadelphia. The search results are shown in the "Number of Retrieved Public Tweets" in Table 2.

Philadelphia's Hotels Occupancy Rate: The data set was provided by Smith Travel Research Inc. (STR), and included average hotel occupancy rates for the hotel industry in Philadelphia, as well as the average hotel occupancy rates of the two groups of selected hotels. The data were collected at both daily and monthly time periods from January 2008 to May 2014.

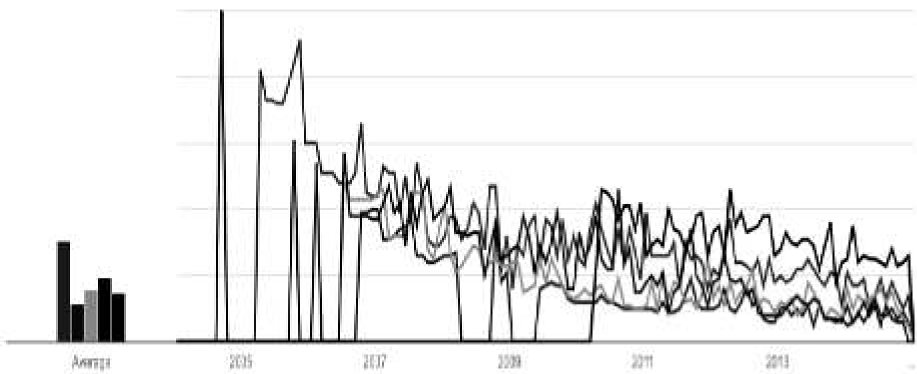
Data Processing

In order to ensure the accuracy and the performance of the analysis, a pre-processing procedure was been carried out on the collected data. Specifically, all the tweets were first transferred to lower case characters, and then any tweet duplications were removed from the data set. In the case of Word-cloud analysis, which will be introduced later, all the hyperlinks, English stop-words (a, the, an, is, etc.), and words generating less value for the analysis, such as “Hotel”, “Philadelphia”, “Philly”, keywords expressing names of hotels (“Four seasons”, “Bellevue”, etc.), included in the tweet postings were also eliminated. Furthermore, since the AKA Ritten house Hotel in Philadelphia does not own a specific Twitter account, the Stay AKA Hotel chain's Twitter account was used to fetch all the posted tweets, and only the tweets containing keywords relating to the hotel in Philadelphia (“Philadelphia”, “Philly”, “Ritten house”, “PA”) was retained for further analysis.

Data Analysis

Trends in the Google Search Volume of Queries Relating to Hotels in Philadelphia: As shown in Figure 1.

Figure 1. Normalized search query volume of 5 different hotels in Philadelphia. From left to right of the legend: Sofitel, Four Seasons, Hyatt Regency, The Latham, and Le Meridien



A chart describing relative comparison of search volume of queries relating to five different hotels in Philadelphia during the year from 2004 to 2014 can be constructed. As observed from Figure 1, there is a decreasing trend in the volume of search queries related to five randomly chosen hotels in Philadelphia over the period from 2004 to 2014. This might result from gradual changes in customer behavior in searching for lodging information. In fact, there are many possible hypotheses supporting this change. For example, as guest experience factors are believed to contribute up to fifty-one percent of the hotel selection decision (Market Metrix, 2010); there might have been a shift from the Google search engine to more industry-focused services like Trip Advisor. Even though there is no sufficiently firm evidence to claim such an alternation, the declining patterns found in the case of hotels in other cities of the United States (e.g. New York) and the increasing number of people consulting hotel reviews at Trip Advisor prior to booking (Anderson, 2012), do support this hypothesis. In addition, the declining variance in the search trend on the Le Meridien hotel, which has operated since 2010, also demonstrates the strong interest of users in a hotel during the very first period after establishment.

Awareness of hotels located in Philadelphia in social media engagement: In this part of the study, various observations on how active the selected hotels are in social media engagement, particularly on the Twitter social network, are tested.

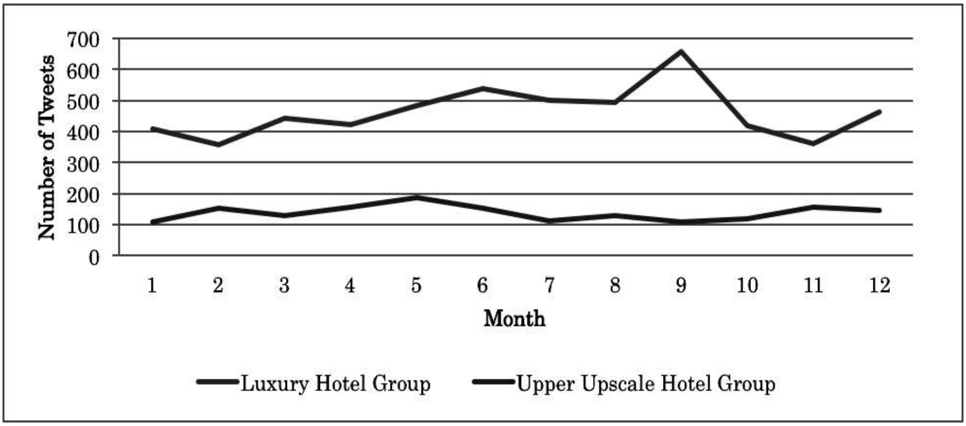
Table 3. Twitter usage statistics for each of the 8 selected hotels

Rank	Group	Hotel	Number of Tweets	Twitter Followers
1	5-star	Four Seasons Hotel Philadelphia	9081	7314
2	5-star	Sofitel Philadelphia	6868	3889
3	5-star International	Kimpton Hotel Palomar Philadelphia	2011	4049
5	5-star International	Windsor Suites	3037	576
4	5-star	AKA Rittenhouse Philadelphia	278*	2761
6	5-star International	Le Meridien Philadelphia	694	308
7	5-star International	Hyatt at The Bellevue	353	440
8	5-star	Latham Hotel Philadelphia	40	91

Table 3 shows the number of tweets and followers of the Twitter accounts for each of the hotels. The rankings given were calculated according to the total number of tweets and followers that each of the accounts has gathered. As we can see, the two 5-star hotels: Sofitel Philadelphia and the Four Seasons Hotel Philadelphia, are ranked as the top hotels that have been exploiting the Twitter service the most, with a total number of tweets of 9081 and 6868 respectively, and the total number of followers of 7314 and 3889 respectively.

The rankings given were calculated according to the total number of tweets and followers that each of the accounts has gathered. As we can see, the two 5-star hotels: Sofitel Philadelphia and the Four Seasons Hotel Philadelphia, are ranked as the top hotels that have been exploiting the Twitter service the most, with a total number of tweets of 9081 and 6868 respectively, and the total number of followers of 7314 and 3889 respectively.

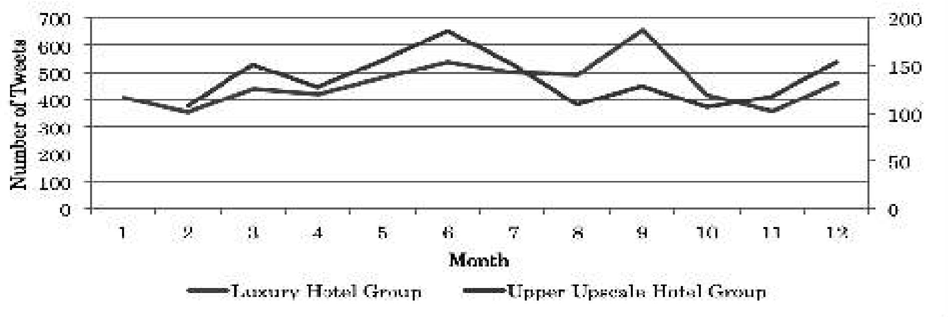
Figure 2. Number of tweets that each of two groups of hotels* posts and re-tweets monthly during the year 2013 - 2014



An interesting observation on the tweeting patterns of the two groups of hotels over the 12 months between 2013 and 2014 is also found when we replace the Windsor Suites hotel with the Hyatt Regency at Penn's Landing, which is classified as a 5-star international hotel. As illustrated in Figure 3, we can clearly recognize that there is big difference in the involvement of the two groups of hotels in Twitter social media over the period. Specifically, the 5-star (luxury) hotel group is 3.5 times more active than the 5-star international (upper upscale) hotel group in terms of the total number of tweets. Moreover, it is also clearly demonstrated that there are also disparities regarding the peak time of posting tweets between the two groups of hotels. To illustrate, the 5-star hotels peak of tweets is around September, while the 5-star international peak is around May.

Interestingly, if we shift the social media usage trend of the 5-star international hotel group forward by 1 month and show it in a secondary axis (see Figure 3),

Figure 3. Number of tweets that each of the two groups of hotels posts and re-tweets monthly during the year 2013 ~ 2014, with the 5-star international group lagged by 1 month.



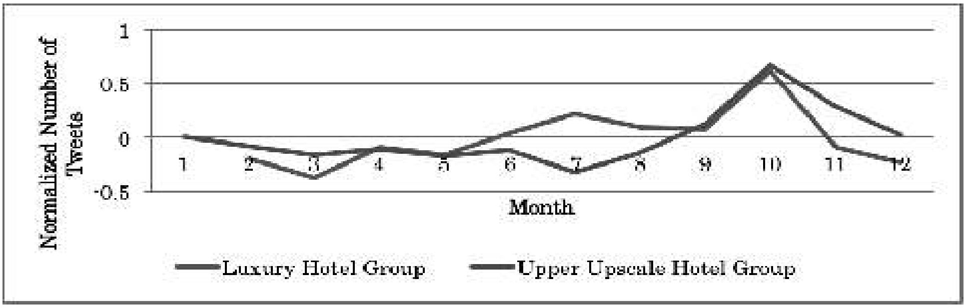
We can see there is a similar pattern in the trends of the two groups of hotels. Which means that the 5-star group of hotels has a similar pattern in social media activities to the the 5-star international group, but 1 month later.

Table 4. AIC result of auto ARIMA analysis in R language on two hotel groups with different additional overlay data parameters: google trend data, management tweet data and Google Trend & management tweets data.

No.	Hotel Group	Data Period	Overlay Data	AIC Result
1	5-star	2013/01 ~ 2014/11	No overlay data	155.46
2			Google Trend data	155.31
3			Tweets data	154.83
4			Google Trend & Tweets data	156.16
5	5-star international	2010/05 ~ 2014/11	No overlay data	331.73
6			Google Trend data	332.69
7			Tweets data	332.95
8			Google Trend & Tweets data	334.17

This might result from differences in the time of commencing lodging service research by two different target consumer segments in relation to the two hotel groups. We subsequently investigated whether this pattern was replicated in the much larger example sets of two groups of hotels. Figure 4 shows the trending patterns of Twitter agagement of a larger set of 5-star and 5-star international of hotels; a total of 5 and 9 hotels, respectively, with the 5-star hotel group tweets being shifted 1 month a head. As observed, the pattern is different from the initial sample set. The two hotel groups share a similar commitment on Twitter during the whole year, except for the period from June to September.

Figure 4. Normalized number of posting on Twitter from a larger set of two groups of hotels[**] during the year 2013 ~ 2014 with the 5-star hotel group lagged by 1 month



Word-Cloud, a qualitative analysis on what's in the tweets: In this analysis, we utilized text-mining techniques to gain insight on the contents of tweets posted by the two groups of hotels. Since the social media data, especially text data such as postings from Twitter or Facebook are usually textual and unstructured data with many noisy words (Barbier & Liu, 2011), it is very challenging to develop criteria to compare the Twitter contents of different hotels. However, we can recognize the textual pattern that those hotels are tweeting online to stay connected with the customers by generating word clouds (Figure 5).

Figure 5. Word cloud of all tweets contents collected from the eight hotels' Twitter accounts

These are an image containing words in a way that the size of each word indicates its frequency or importance. By doing this, we can understand the differences in the actual usage of the Twitter social network of two groups of hotels, and even among the hotels.

In addition, we also recognize some interesting patterns in the word-cloud of hotel Sofitel Philadelphia, and the Bellevue at Philadelphia. Regarding the Sofitel hotel, the word “concierge choice” is one of the top used words in tweets, and might imply that the hotel attracts customers by regularly updating recommendations of interesting places (e.g. opening museum, famous attractions), restaurants, or events (e.g. new movies, festivals, etc.) around the hotel area. This information is usually imparted by the concierge in a major hotel. With regard to the Bellevue at Philadelphia hotel, since the word “Wedding” and “Photo” appear a lot in the hotel tweets, we can say that the hotel pays a lot attention to advertising its wedding services through Twitter.

The role of management tweets and google trends in predicting hotel occupancy: in this section, timeseries forecasting analysis was carried out to examine whether Google trend data and management tweets (tweets posted by hotel managers) support the predictions of the hotel occupancy rates in the two hotel groups studied. In order to achieve this, we carried out Auto-Regressive Integrated Moving Average (ARIMA) timeseries analyses in R language (auto.arima function) for each of the hotel groups, using four overlay data options. Akaike's Information Criterion (AIC) was used to determine whether management tweets and Google trends data contribute to a better hotel occupancy prediction model. Moreover, because of inconsistencies in the number of collected tweets among the hotels, an appropriate data period had to be chosen, so that all of the tweets posted in that period were available across all of the hotels in the target hotel group.

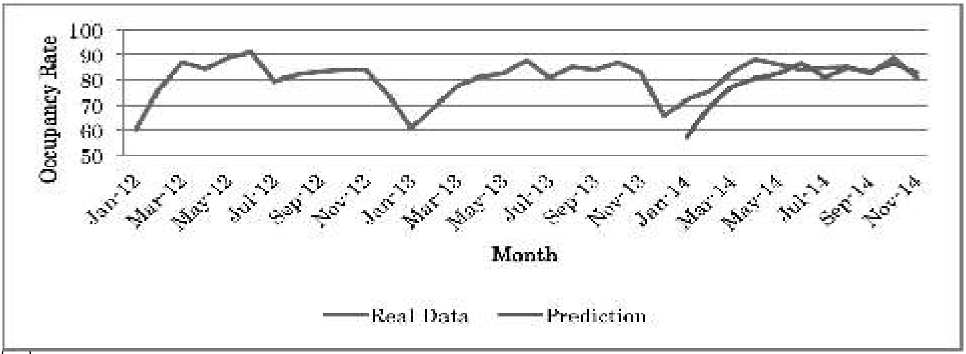
Even though the AIC index (the smaller the better the predictability of model parameters) is not used to compare the performance of each different model, AIC can be used to compare the performance of a model under different parameters. The results show that Google trend and management tweet data did not contribute in improving the predictability of the base ARIMA model during the examining time period. Nevertheless, we continued to carry out Arima analysis on the data relating to the 5-star international group of hotels for a different time period: May 2010 ~ December 2013 (Table 5), and sought to forecast the occupancy rate of the 5-star international group of hotel in the year 2014.

Table 5. AIC result of auto ARIMA analysis in R language on the 5-star international hotel group with overlay data initiated in the period from May 2010 to December 2013

No	Hotel Group	Data Period	Overlay Data	AIC Result
1	5-star International	2010/05 ~ 2013/12	No overlay data	266.64
2			Google Trend data	267.32
3			Tweets data	268.61
4			Google Trend & Tweets data	122.97

The result gave a better AIC index for the Arima model, with the additional overlay data of both Google trend and management tweets, in comparison with other parameter options. This model was then tested to forecast the hotel occupancy rate in the time period from January 2014 to November 2014. The predicted occupancy rate of the 5-star international group of hotels is shown in Figure 6.

Figure 6. Occupancy prediction for the 5-star international hotel group using the Arima model with both Google trend and management tweets data as overlay data parameters



The calculated Mean Square Error of the prediction period (January 2014 ~ November 2014) is approximately 1.98, which is a fairly acceptable result in terms of the predictability of occupancy rates from social media data.

Limitations and Future Research : There are a number of limitations in this study, which are also promising future researching directions. Even though the study has carried out analysis regarding how the engagement of two hotel groups is in relation to Twitter social networking; it could not recognize whether the findings on the engagement pattern are right for all the hotels studied. This might have resulted from the limitations in the collection of management tweets from every hotel. Moreover, because the generated word-cloud depends on not only the number of words, but also on the number of collected tweets (which are largely different among hotels), other text-mining analyses should also be done to further support this research. Furthermore, the Arima anlysis with overlay data of different timeline periods can be modified to draw a more robust conclusion. Additionally, other types of social media data, such as hotel reviews, facebook posts, etc. can be furthered considered as promising predictors in the task of forecasting hotel occupancy.

Future work will further address questions of reliability and validity of the sentiment measuring instrument advanced here. Further work could be performed to correlate external measures of reliability and validity with the derived sentiment measures. This could be achieved by monitoring the sentiment curve in real time and administering online questionnaires periodically. Measures of correlation between the scores measured in the questionnaires and the sentiment curve could then be calculated. Akehurst (2009), noted that "the time, energy, resources and costs required to locate relevant user generated content and extract useful and meaningful information was currently too much at the present time", however we believe that data mining applied to social media can explain and reflect both numerically and visually societal trends, and in particular, those in the tourism industry.

Conclusion

This paper describes the different approaches in analyzing social media data, but concentrated on Google trend and Twitter postings data specifically. Quantitative analysis on the time-series, and text-mining analyses on the tweet contents revealed how competitive the two groups of hotels are at engaging in the Twitter social network. Despite the weak support of management tweet and Google trend data in predicting hotel occupancy using a time-series forecast technique; there appear to be certain time periods in which those social media data play a more contributive role. Therefore, more research investigating the correlation relationships between Google trend, management tweets and public tweets data is very promising in generating a more reliable result in the near future.

The literature and practical experience in the tourism industry emphasizes that organizations depend on up-to-date and relevant market knowledge to compete in facility provision, and to promote the industry, as well as to manage the relationship with tourists. This paper shows that new forms of social media provide valuable and previously difficult to obtain real-time knowledge on tourist perceptions, concerns, and sentiment towards accommodation providers, and that these can be generalized. We have shown how analysis of comments from such social media as Twitter micro-blogs can be used to reveal potential and recent tourists motivations and travel sentiment relating to high quality hotels. The data gathering, preprocessing, analysis, and sentiment mining methodologies developed and used in this study supply tangible, actionable, and beneficial knowledge for predicting tourist movements and analyses of the tourism industry in any location (Zheng & Pan, 2011).

Sentiment mining is beginning to be applied in the context of tourism and hospitality, and demonstrates an ability to examine large data sets for trends that can predict what consumers will choose in terms of a tourism destination, and also which areas they will avoid for emotional reasons. Such analyses imply that further work in this area using alternate data sets similar to those we have utilized in this study may provide tourism researchers with a rich source of knowledge for transfer to operators and the tourist. When the methodology we developed is focused on actual locations it will provide knowledge of market conditions and ways to move information to the place or recipient where it is most required.

References

Adam, J., Cobos, X., and Liu, S. (2007). *Travel 2.0: Trends in industry awareness and adoption*. New York: New York University and PhoCusWright Inc.

Akehurst, G. (2009). User generated content: the use of blogs for tourism organizations and tourism consumers. Service Business 3(1): 51-61.

Anderson, C. K. (2012). The impact of social media on lodging performance. *Cornell Hospitality Report 12*(15): 4-11.

Barbier, G., & Liu, H. (2011). Data mining in social media. *Social Network Data Analytics*, 2011: 327–352

Beaumont, C. (2008). *Tweet, tweet, here comes Twitter*. http://www.telegraph.co.uk/technology/3357007/Tweet-tweet-here-comes-Twitter.html. Accessed 2 February 2015.

Blal, I., and Sturman, M. C. (2014). The Differential Effects of the Quality and Quantity of Online Reviews on Hotel Room Sales. Cornell Hospitality Quarterly, May 6. Online: http://cqx.sagepub.com/content/early/2014/05/05/1938965514533419.abstract?rss=1. Accessed 2 February 2015.

Carson, D. (2005). An overview of developing regional tourism using information communications technology. In Marshall, S., Taylor, W., and Yu, X. (Eds), *Encyclopedia of Developing Regional Communities with Information and Communication Technology*. Idea Group, Hershey, PA.

Choi, H and Liu, Paul (2011). Reading Tea Leaves in the Tourism Industry: A Case Study in the Gulf Oil Spill. Google Research Blog, March 24. http://googleresearch.blogspot.jp/2011/03/reading-tea-leaves-in-tourism-industry.html. Accessed 2 February 2915.

Choi, S., Lehto, X. Y., and Morrison, A. M. (2007). Destination image representation on the web: content analysis of Macao travel related websites. *Tourism Management 28*(1): 118-129.

Claster, W. B., Pardo, P. D., Cooper, M., and Tajeddini. K. (2013). Tourism, travel and tweets: algorithmic text analysis methodologies in tourism. *Middle East Journal of Management*, *1*(1): 81-99.

Claster, W. B., Cooper, M., Isoda, Y., and Sallis, P. (2010). Thailand -- Tourism and Conflict: Modeling Tourism Sentiment from Twitter Tweets Using Naïve Bayes and Unsupervised Artificial Neural Nets. *CIMSim2010. Computational Intelligence, Modelling and Simulation*, 89-94.

ComScore. (2007). *Online consumer-generated reviews have significant impact on offline purchase*, November 29. http://www.comscore.com/press/release.asp?press=1928. Accessed 9 December 2014.

Dellarocas, C. (2003). The digitization of word of mouth: promise and challenges of online feedback mechanisms. *Management Science 49*(10): 1407-1424.

Dey, L., Haque, Sk M., Khurdiya, A., and Shroff, G. (2011). Acquiring competitive intelligence from social media. In Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data. ACM.

Ellion Travel Technology Group. (2007). *Web 2.0 and the Travel Industry: Practical strategies for exploiting the social media revolution*.

http://www.ellion.co.uk/sectors/travel/index.php. Accessed 20 April 2012.

e Marketer. (2007). *UK Online Travel: Travelers are sharing experiences online*, 25 June. http://www.emarketer.com/Article.aspx?id=1005067andsrc=article2_newsltr. Accessed 20 April 2014.
Forrester Research Group. (2007). *Social Techno graphics*.
http://www.forrester.com/rb/Research/social_technographics%C2%AE/q/id/42057/t/2. Accessed 20 April 2014.

Go, A., Huang, L., and Bhayan, R. (2009). *Twitter Sentiment Analysis*. www.nlp.stanford.edu,2009. Accessed 20 April 2014.

Hepburn, C. (2007). *Web 2.0 for the tourism and travel industry: A white paper*. http://blogs.reactive.com/RequestWhitepaper.aspx. Accessed 20 April 2014.
Horrigan, J. (2008). *Pew Internet & American Life Project*, February 13. *Pew Research Center's Internet & American Life Project*.

http://www.pewinternet.org/Reports/2008/Online-Shopping/01-Summary-of-Findings.aspx. Accessed 20 April 2014.

Hu, M., and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '04): 168-177.

Jansen, B. J., Zhang, M, Sobel, K, and Chowdury, A (2009). *Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Sciences and Technology 60*(11): 2169–2188.

Jindal, N., and Liu, B. (2006). Identifying comparative sentences in text documents. *ACM Digital Library*. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '06): 244-251.

Kennedy, A., and Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence 22*: 110-125.

Laboy, F., and Torchio, P. (2007). *Web 2.0 for the travel marketer and consumer: a white paper*. http://www.sabrehospitality.com/pdf/whitepaper.pdf. Accessed 10 January 2015.

Lanz, L. H., Fishhof, B. W., Lee, R. How Are Hotels Embracing Social Media in 2010 – Examples of How to Begin Engaging. New York: HVS Sales and Marketing Services. www.hvs.com/staticcontent/library/nyu**2010**/.../SocialMediaIn**2010**.pdf. Accessed 20 April 2014.

Pang, B., and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval 2*(1-2): 1-135.

Rabanser, U., and Ricci, F. (2005). Recommender systems: do they have a viable business model in e-tourism. In *Proceedings of the ENTER Conference, Innsbruck, Austria, 2005*: 160-171.

Sarks, G. (2007). *Consumer generated content in travel: A roadmap for bringing consumers into the conversation*. Web-seminar presentation, 16 April. New York: Compete, Inc.

Senecal, S., and Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of Retailing 80*: 159-169.

Werthner, H., and Ricci, F. (2004). E-commerce and tourism. *Communications of the ACM – The Blogosphere 47*(12): 101-105.

He, W., Zha, S., and Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry, *International Journal of Information Management*, *33*(3): 464-472.

Zhuang, L., Jing, F., and Zhu, X. (2006). Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management* (CIKM '06): 43-50.

Zheng X., Pan, B. (2011). Travel queries on cities in the United States: Implications for search engine marketing for tourist destinations. *Tourism Management*, *32*(1): 88-97.

About the Authors

Malcolm J. M. Cooper PhD LLM is Emeritus Professor, Graduate School of Asia Pacific Studies, Ritsumeikan Asia Pacific University (APU), Beppu, Japan. He teaches tourism management and environmental law. Formerly Vice President for Research and International Affairs, he is a specialist in tourism resource management (natural, cultural, social) and tourism development, environmental and water resource management and environmental law, and has worked and published widely in these fields.

Kazem Vafadari PhD is Associate Professor, College of Asia Pacific Studies, APU, Beppu, Japan. He teaches tourism policy and planning, tourism research, and cultural heritage management. He is an expert on tourism and developing countries in Asia, and on agricultural heritage landscapes and rural development, and works with the United Nations Food and Agriculture Organization as an advisor and scientific committee member for the Globally Important Agricultural Heritage Systems (GIAHS) project.

William (Bill) Claster PhD is Professor, College of Asia Pacific Studies, APU, Beppu, Japan. He teaches Data Mining and Knowledge Management, Maths, Statistics, Computing and Physics.

Phillip D. Pardo MBA is Professor of Accounting and Business Analytics, Graduate School of Management, APU. His research work covers Innovation in the Accounting and Hospitality sectors as well as the effects on Management of CGM Reviews, Sentiment

Mining and Customer Relationship Management (CRM); recent publications have focused on Medical Tourism and Social Media Analytics and Sentiment Mining in the Tourism and Hospitality industry.

Le Quang THAI was born in Khanh Hoa Province, Vietnam. He is presently pursuing his studies at Ritsumeikan Asia Pacific University (APU) in the field of Business Administration focusing on Strategic Management. His research interests include applied programming in business, modeling for decision-making and machine learning. He has been an IEEE member since 2014.