Article

A Thematic Travel Recommendation System Using an Augmented Big Data Analytical Model

Suriya Priya R. Asaithambi [1], Ramanathan Venkatraman [1] and Sitalakshmi Venkatraman [2,*]

[1] Institute of Systems Science, National University of Singapore, Singapore 119615, Singapore
[2] Department of Information Technology, Melbourne Polytechnic, Preston, VIC 3072, Australia
* Correspondence: sitavenkat@melbournepolytechnic.edu.au; Tel.: +61-3-9269-1711

Abstract: Tour planning has become both challenging and time-consuming due to the huge amount of information available online and the variety of options to choose from. This is more so as each traveler has unique set of interests and location preferences in addition to other tour-based constraints such as vaccination status and pandemic travel restrictions. Several travel planning companies and agencies have emerged with more sophisticated online services to capitalize on global tourism effectively by using technology for making suitable recommendations to travel seekers. However, such systems predominantly adopt a destination-based recommendation approach and often come as bundled packages with limited customization options for incorporating each traveler's preferences. To address these limitations, "thematic travel planning" has emerged as a recent alternative with researchers adopting text-based data mining for achieving value-added online tourism services. Understanding the need for a more holistic theme approach in this domain, our aim is to propose an augmented model to integrate analytics of a variety of big data (both static and dynamic). Our unique inclusive model covers text mining and data mining of destination images, reviews on tourist activities, weather forecasts, and recent events via social media for generating more user-centric and location-based thematic recommendations efficiently. In this paper, we describe an implementation of our proposed inclusive hybrid recommendation model that uses data of multimodal ranking of user preferences. Furthermore, in this study, we present an experimental evaluation of our model's effectiveness. We present the details of our improvised model that employs various statistical and machine learning techniques on existing data available online, such as travel forums and social media reviews in order to arrive at the most relevant and suitable travel recommendations. Our hybrid recommender built using various Spark models such as naïve Bayes classifier, trigonometric functions, deep learning convolutional neural network (CNN), time series, and NLP with sentiment scores using AFINN (sentiment analysis developed by Finn Årup Nielsen) shows promising results in the directions of benefit for an individual model's complementary advantages. Overall, our proposed hybrid recommendation algorithm serves as an active learner of user preferences and ranking by collecting explicit information via the system and uses such rich information to make personalized augmented recommendations according to the unique preferences of travelers.

Keywords: recommendation system; big data analytics; thematic recommendations; tourism

1. Introduction

In data-driven markets, digital platform competitors reliably offer effective advice. Recommendation engines are a powerful tool for internet companies [1,2]. While thematic recommendation systems employ deep technologies such as deep and reinforced learning for data streaming in eCommerce applications (platforms such as Spotify, Netflix, and Amazon), other sophisticated sectors such as travel, insurance, logistics, and healthcare are still in their early days of adoption. There is a need for smart recommendation systems [3,4]. Particularly in the travel sector, travel planning companies that started in the past two decades (e.g., Skiplagged in 2013, Google Flight in 2011, Airbnb in 2008, Kayak in 2005,

HomeAway in 2004, SkyScanner in 2003, TripAdvisor in 2000, and Expedia in 1996) could have sustained their services with smart recommendation systems.

Currently, travel decisions are predominantly made without the full knowledge of several alternatives and options based on user-specific preferences. Potential travelers resort to travel agents to recommend rather than make their own decisions due to a lack of information/awareness of new destinations. Furthermore, the information available over the Internet adapted with filtering and selection methods used by search engines can provide limited guidance only. With large amounts of information available online, travelers can get overwhelmed during their decision-making process. With recent developments in technologies and practical tools available to process big data, there exists an abundant opportunity to build automatic or semiautomatic systems that would provide a range of alternatives and suggestions based on user preferences to help travelers in their decision-making process. These form the key motivations of this research work that aims to design and develop a prototype of a travel recommendation system with edge technologies and effective models. However, prior to scoping an effective travel recommendation system, we considered the current state-of-the-art practices in the industry and their limitations.

At present, most of the key travel industry players continue to rely on a destination-based recommendation approach in the form of travel packages [5]. However, more inclusive data-driven insights are required for modern travel recommendation systems to provide personalized advice. A smart recommendation system could be developed to meet the varying needs of individual tourists in terms of personal tastes, interests, travel habits, and other contextual preferences influenced by time and monetary constraints [6,7]. Currently, potential tourists are provided with a plethora of choices from a variety of fixed package-based recommendations that can lead to decision fatigue; furthermore, they fail to meet the consumer requirements effectively [5,8].

Recommender systems form a subclass of information filtering system to predict users' preferences and recommend items that are likely interesting to them. Companies using recommender systems focus on increasing sales as a result of very personalized offers and an enhanced customer experience [7]. Recommendations typically speed up searches and make it easier for users to access the content they are interested in, and then surprise them with diverse offers they would have never searched for. Smart recommendations depend on intelligent modeling of the information filtering process, which is a key focus of this research work in developing a travel recommendation system that has a potential contribution to the growth of the tourism industry.

Tourism is a significant part of many national economies, and the immense shock to this sector resulting from the pandemic is affecting the wider economy [9]. Travel recommendations are considered one of the most complex tasks, as they use a variety of data for models, such as images, textual descriptions, user trip details, user ratings, and information influenced by other users. While many travel recommender systems are data-centric, they lack user focus and situational or impulsive aspects that play a major role in the decision making of potential tourists. Several dynamic changes in the global environment such as weather, political situation, natural disaster, pandemic status, and news feeds are not considered. The lack of new-generation recommender systems capable of considering such big data to include context-, user-, and sentiment-aware analytics forms the key motivation for this paper. To address this gap in the literature, we focus on proposing a user-centric context-aware model by employing suitable augmented big data analytics for developing a prototype travel recommender system (TRS). Furthermore, the travel and tourism industry desires to grow from a traditional tourist destination image (TDI)-based recommendation services to a more accommodative smart and interactive solution. This forms the focus of our approach in proposing a smart travel recommendation system in this research work.

In this paper, we propose an augmented model to integrate analytics of a variety of big data (both static and dynamic) such as destination images, reviews on tourist activities, weather forecasts, and recent events via social media for generating more user-centric

and location-based thematic recommendations efficiently. This would address the current limitations in online tourism services and enhance existing recommendation models that predominantly use text mining for "thematic travel planning".

We consider two key aspects critical in our approach toward developing an effective travel recommendation system. Firstly, our proposed system should be able to manage data that are dynamic. Dynamic data comprise varied scaling, freshness, and stringent factors against noise. Secondly, our recommender system should be contemporary in adopting a hybrid processing model that can be divided into parts: front-end, transport, storage, serving, online training model, and offline training model. Inclusiveness comes from domain expertise in identifying untapped data resources that are either internal or external to the organization. For example, with the evolution of social media and communication technologies, online travel reviews (OTRs) have grown significantly, allowing users to post their travel experiences, opinions, comments, and ratings in a structured way. Recent studies have proposed models to analyze information displayed in search engines for measuring TDI but have not considered OTRs that are not indexed in search engines [10]. Furthermore, with the advent of behavioral analytics, personalized recommendations proliferate the space of eCommerce, music, social media, and TV. This presents a significant opportunity to enhance the capabilities and offerings of traditional travel recommender systems that currently exist predominantly on the basis of location information and social contexts [11,12].

The prototype of our thematic dynamic travel recommender system employs prediction models with four categories of data collected, namely, (i) images, (ii) reviews, (iii) climate, and (iv) social media, along with location-based user constraints and preferences. The value proposition of our proposed model works on various levels of personalization such as generic, targeted, transient, and persistent. Our travel recommendation algorithm also considers popular destinations, users' personal preferences, recent travel history, and long-term travel preferences. Thus, the recommendation prototype caters to both individual and travel agent needs. On the basis of the initial results, the recommendation system widens the search of travel destination options to second-tier cities. Furthermore, our experimental study revealed that, when the user articulates specific preferences similar to search engines, the recommendation converges more quickly.

Our proposal leverages using augmented big data, along with collaborative filtering techniques and crowdsourcing, to build an effective recommendation system that assists the user in their decision making for tourism activities. We used a big data processing architecture capable of storing, consolidating, and spotting new trends and insights from disparate information associated with various travel destinations and user contexts. Our implementation first consolidates the heterogeneous data from various sources, filters the relevant information, and creates semi-structured data frames using a scalable framework *(Apache Spark)*. The mix of structured and unstructured data is then preprocessed and analytically modeled using machine learning techniques such as time series forecasting, sentiment mining, and artificial neural networks. The new innovative hybrid model generates insights that are inclusive of users' personal preferences *(highest priority)*, environmental factors, and destination-based features. These analytical insights form the basis of our front-end information retrieval app (prototype implementation) that provides destination suggestions according to current user preferences. Furthermore, our TRS provides a user-friendly interface, which allows users to articulate their requirements and preferences using free text that makes the interaction more conversational. One tricky challenge was to filter essentials from the massive unstructured news feed that were relevant to potential destination choices. While the current model can pick hints from the bag of words it is being trained on, training and fine-tuning can be improved in further extension of our work with more data from the internet.

Overall, this paper is aimed at developing an end-to-end prototype TRS to evaluate our proposed big data-based recommendation model using augmented user-centric data analytics of various themes and arguments with relevant current environment details of

potential destinations. Since the model was built for cluster deployment and uses a general scalable framework for analytics modeling, automatic scaling and model tuning are flexible in case of future extensions and fine-tuning.

The remainder of the paper is structured as follows: Section 2 presents the background of this study by providing a detailed review of the literature with a comparative analysis of existing models of TRS and establishes the relevance of big data in developing content-based TRS for the future. We explore various big data techniques for the creation of such a content-based TRS. In Section 3, we study the state-of-the-art recommendation techniques adopted by existing TRS solutions in both commercial systems and academic research studies. This section focuses specifically on identifying the merits and limitations of existing models to propose an enhanced TRS model. In Section 4, we describe our proposed thematic TRS using an augmented big data analytics model and prototype development. Section 5 provides the results obtained with the evaluation of our proposed TRS prototype using a sample user-centric and context-based dataset. Lastly, Section 6 concludes with a brief description of the outcomes achieved and suggestions for future research.

2. Background of the Study

With increasing data availability in online travel systems, the challenge of assisting users with an efficient way to narrow down their search in arriving at a suitable travel option has grown dramatically. Big data are dynamic, unpredictable, and largely consumer-driven. It is the direct consequence of ubiquitous technology and Internet connectivity available all the time. With the advancement in big data processing, recommendation engines have sharpened their ability toward better predictions. Although there are fewer changes in the foundational computation models, the newer processing techniques, collaborative filters, and behavioral aspects of users could influence the way in which recommender systems are built over time. This section systematically presents various commonly adopted recommendation models, the evolution of three generations of recommendation engines, a comparative analysis of the literature on novel works, and the relevance of big data in developing an effective travel recommendation system.

2.1. Travel Recommendation Modeling

Commercial travel intermediaries (e.g., Expedia.com or TripAdvisor.com), tourism agencies, and government organizations in various destinations employ TRS to increase turnover (e.g., sell more hotel rooms and increase advertising revenue). According to studies regarding the models of TRS, there are six different classes reported in the literature [1]. We describe the commonly adopted recommendation models below.

Content-based: This model learns to recommend items that have similar features in the information content as compared to those that the customer liked or rated in the past. For example, in Booking.com, if a customer has booked a hotel with the feature "pet-friendly" in their information content, the system learns to recommend other hotels with this feature for future bookings.

Collaborative filtering: This model recommends to the customer items that other users with similar tastes have liked in the past. This is sometimes referred to as people-to-people correlation and uses neighborhood-based methods. Users are recommended to destinations depending on similar behavior pattern of other users collaboratively.

Demographic: This model recommends items according to the demographic profile of the customer, such as gender and age, and it may not need the history of the customer. For example, users are directed to specific travel web sites according to their language or country. However, earlier studies showed that using demographic information alone is not sufficient for an accurate prediction of ratings in recommending the most relevant attractions to the tourists [5].

Knowledge-based: This model recommends items according to specific domain knowledge about how certain item features meet customer needs and preferences, and how the items are useful to the customer based on conversational dialogue. For instance, a destina-

tion management organization can advertise a bundle or complementary attractions to a specific customer according to the information collected such as destination, itinerary, and purpose of travel.

Community-based: This model recommends items according to the preferences of the friends of the users. For instance, TripAdvisor learns to recommend hotels that have been rated positively by the users' friends on Facebook.

Hybrid recommender systems: These systems are based on the combination of the abovementioned models and uses the technique that best suits the situation. For instance, collaborative filtering methods suffer from new-item problems, making them unusable to recommend items that have no prior ratings. This is not an issue for content-based approaches since the prediction for new items is based on their given features. Given several approaches, hybrid systems can provide more contextualized and personalized recommendations to a customer, e.g., different recommendations for winter versus summer vacation.

2.2. Evolution of Travel Recommendation Systems

Recommendation systems can be classified into three generations on the basis of the technology maturity they demonstrate, end-user involvement necessitated, and optimal results provided from the traveler's perspective [13,14]. The first-generation recommendation systems were from expert traveler opinions, according to top attractions, rankings, and opinion blogs made by other travelers. This pioneer generation does not incorporate a personalization feature. The second-generation travel recommenders are aimed at reducing the explicit user inputs required, which results in reduced overhead for the traveler. At this stage, personalization is achieved through classification and prediction processes. The third-generation travel recommender systems attempt to incorporate multidimensional aspects for the recommendations. They attempt to provide solutions to all associated activities that a traveler needs to consider during planning.

2.3. Comparative Analysis of the Literature

With the increasing utilization of location sharing services and applications, location-based social networks (LBSNs) have enlarged the scope of TRS to include geospatial location and location-related content. With advancements in web/mobile technologies and social media, novel approaches with the use of contemporary user features and data attributes are emerging in modeling TRS. Some of the recent developments summarized in these studies suggest that the use of sentimental attributes along with attributes for point of interest (POI) mining can improve the recommendation performance. However, problems such as cold start continue to be challenges that need to be addressed and solved effectively [15]. Whereas eCommerce sites mostly apply knowledge-based and content-based approaches, and recommendations are confined at the destination level [7], recent research studies have adopted collaborative filtering more commonly in TRS [8,9]. In Table 1, we identify the merits and demerits of novel studies on travel recommender systems covered by the surveys in [6,13,14].

2.4. Big Data Relevance for Travel Recommendations

At a tactical level, big data refer to "datasets which could not be captured, managed and processed by general computers within an acceptable scope" [16]. At a more strategic level, big data represent the "information assets characterized by such a high volume, velocity, and variety to require specific technology and analytical methods for their transformation into value" [17]. Travel industry players such as travel suppliers, online travel agencies, and global distribution systems can access an extensive amount of data that are captured during normal business interactions across the travel value chain [18]. Furthermore, players from the closely related industries, such as the airline, hotel, and food and beverage industries, also hold rich information. While these information sources cannot be mapped at a user level, an accurate augmentation of core travel-related data is possible

on the basis of a destination or a specific tourist attraction. Lastly, another significant proportion of travel-related information can be obtained from external sources such as travel blogs, online travel reviews, and social media posts as user-generated content. Such content can take different forms including texts, images, or videos.

Table 1. A comparative analysis of novel studies on travel recommender systems.

SN	Title	Merits	Demerits
1	A novel POI location-based recommender system utilizing user location and web interactions	Sparsity problem is solved	Accuracy problem exists
2	A rich sentimental travel schedule via sentimental POI mining and recommendation	Sentiment analysis is used for POI mining	The sparsity problem is not examined
3	Author topic model-based collaborative filtering for personalized POI recommendations	Sparsity problem is solved	Only textual information is used
4	User-based collaborative filtering for tourist attraction recommendations	Efficient in tourist attraction recommendation	Number of visits is considered as the implicit feedback
5	A personalized geographic-based diffusion model for location recommendations in LBSN	Nearby POIs can be recommended using a single model	Time of check-in and categories of location are not considered
6	An HITS-based POI recommendation algorithm for LBSN	Performs well in precision and recall	Geographical distance is not considered
7	A random walk around the city: new venue recommendation in location-based social networks	Predicts previously unvisited places	Not efficient for abundant data
8	Price and profit awareness in recommender systems.	Incorporates purchase-oriented information into recommendation algorithms	Only short-term perspectives are considered
9	Semantic wordnet-based approach	Strong conceptualization and connectedness	Interpretations are weak and subjective
10	Heuristic search for recommendation	Extra information about mode of transport	Cannot effectively recommend points of interest in a new city
11	Fuzzy logic and use of genetic algorithm	A detailed itinerary with time duration for each point of interest	Not completely accurate recommendations
12	Hybridization and machine learning	High accuracy	Suggestion is only based on points of interest; nothing considered for tour plans

As a result, travel and transportation CEOs ranked the "information explosion" due to big data as a key reason for the transformation of their organizations [19]. Such transformation is visible from the reduction in the number of travel agents in the US from 124,000 in 2000 to 74,000 in 2014, as reported by Atlantic [20]. While this signifies a gradual shift and growth of virtual travel agents, the pace of adaptation seems quite slow when compared to other related industries such as airlines or taxi services for local commuting. A key reason behind this is the unavailability of holistic and intelligent analytical frameworks that can provide customized recommendations taking a user's requirements into consideration, similarly to how a travel agent does so traditionally [21]. As demonstrated in Section 3, the

available commercial solutions cover only some of the big data aspects such as volume and velocity, thereby limiting the potential of such solutions. To have a broader context about an optimal solution, we first evaluate different travel-related data sources as a function of the seven key characteristics of big data, as summarized below.

Volume: Volume covers data from multiple sources, e.g., travel blogs, online travel review websites, and social media in the form of travel experiences. Another source is the mobile device location-based sensor network. Studies have found much value in using sensor-based information data such as geolocation to push real-time marketing promotions to tourists [22,23].

Velocity: Velocity in the context of tourism data can be interpreted in two ways, i.e., the speed of input data generation and the speed of responsiveness [24] The former includes capture, storage, and analysis [1,25]. In general, a little loss in velocity for improved business value is acceptable in practical situations [1,25].

Variety: Variety refers to all the structured and unstructured travel data collected manually and automatically by systems, which can be used for a better customer experience when analyzed using advanced analytical and machine learning techniques [10,24,26].

Veracity: Veracity relates to the accuracy and truthfulness of acquired information, in the context of emerging problems, such as fake and paid reviews that now impact approximately 15% of the data. In this context, data entry errors by customer service executives can play a major role at this stage. This is further emphasized by a low 54% trust coefficient of online reviews, although user-generated content that is known to be unsolicited and unbiased can contribute toward the accuracy of a recommendation system [2,10].

Value: The value of all these sources of information can be described by their importance in driving better and more tailored recommendations for the users. Destination images have also been used in different ways, either to capture a customer's lifestyle attributes or to detect their choices and preferences [24]. Twitter data have also been used for the same purpose in other studies [18]. Weather and cost information by actual users can drive higher reliability of the findings [25].

Connectedness and Infer-ability: Although external data are only loosely connected at a destination or tourist attraction level, they do not impact the infer-ability aspects when aggregated over many users [10]. Furthermore, the infer-ability is positively impacted by the unsolicited content generation process, without any bias or modification in the absence of any explicit incentive commitments. On the negative side, infer-ability is affected by the absence of proper feedback from users about the importance of specific information in their decision-making process. Table 2 summarizes the criticality of these different big data aspects for tourism data.

Table 2. Importance of different big data aspects for tourism data.

Volume	Velocity	Variety	Veracity	Value	Connectedness	Infer-ability
High	Low	High	Medium	High	Low	Medium

Among the seven big data characteristics, volume, velocity, and variety are the top three aspects that have attracted the IT professionals to delve into big data capabilities for tourism industry [19]. The concepts of big data have been borrowed and leveraged in multiple ways in this domain, ranging from the most basic premise of handling massive datasets to integrating disparate data types for the generation of real-time insights [23]. The real-time applications of big data concepts have, however, largely focused on tackling transactional issues of travel, such as route planning [19], pushing location-based marketing promotions [23], finding cheapest tickets, or generating alerts pertaining to next best offers [20].

More strategic applications of big data techniques and infrastructure focus primarily on the supply side for driving better decisions about market segmentation, spend behavior analysis, and return-on-investment optimization [22]. Using the same supply side of infor-

mation, different studies have focused on specific areas and usages of big data, including the creation of aggregated dashboards [27], dynamic product pricing, and prediction of tourist volumes [25]. However, the key limitation of any such analysis is that it completely overlooks the user-related information and ends up recommending the same or similar options to all users. In [23], the author argued and articulated concisely that the "next best action is different to the next best offer, an approach that prevails today". Following this train of thought, the author emphasized the relevance and importance of the demand side of information that originates from a user's preferences and choices. Integration of demand levels can facilitate the revenue maximization process by improving the conversion rates. This is due to the tailoring of the management process for service offerings with the foremost requirement of meeting the customer's preferences.

Recently, several studies have come up with interesting findings using such user-generated information. In [27], the authors used a centralized repository for user metadata, including the user's search history, social media features, and marketing campaign-related information to create four different indices related to the customer, namely, visit, health, wealth, and lifestyle. While the approach was objective and user-focused, it assumed that the quasi-static user features were sufficient to infer about travel decisions and, in turn, ignored the importance of the situational or impulsive aspects. Potentially, this can lead to recommendations of the same or similar places to a user every time, reducing the variety of options with a user's evolving taste. This issue can effectively be dealt with by overlaying the framework proposed in [23], which suggested combining a natural long-term propensity of purchase of a customer to a short-term intent as captured by spontaneous signals such as web searches. Similarly, a more advanced study aimed at the creation of a "destination image" that captures the traveler's perception about a place by aggregating a large corpus of travel reviews [10].

Given these encouraging findings in recent research studies, we performed a detailed review of the existing commercial TRS solutions, identifying their strengths and limitations with respect to the big data opportunities identified above. We also delved deeper into the analytical aspects to help create an objective proposal for our prototype solution development.

3. Analysis of Current Techniques

3.1. TRS Interfaces and Functionalities

It is common to have TRS interfaces presented with a web and/or mobile device orientation. Due to the exponential increase in mobile devices, most web-based TRS also have a mobile device counterpart.

- Web-based recommenders are highly user-friendly with easy access to information from other related sources, such as maps, images, and videos. Examples include City Trip Planner [28], e-Tourism [29], Otium [30], and EnoSigTur [31].
- Mobile recommenders are more focused on providing only the relevant and essential information and are designed to be portable for travel [32]. Examples include MapMobyRek [33], MoreTourism [34], and LiveCities [35].
- According to [36], the two most successful web-based recommender system technologies are Triple hop's TripMatcher and VacationCoach's expert advice platform, Me-Print (used by travelocity.com).
- TripleHop's TripMatcher is a recommendation software based on artificial intelligence and human knowledge that advises users on the destinations that best match their needs and preferences [37].
- VacationCoach exploits user profiling by explicitly asking the user to select one from a set of predefined traveler profiles, which induces implicit needs that the user does not provide. Alternatively, the user may choose to provide precise profile information by completing the appropriate data entry form [38].

Both recommender systems try to mimic the traditional offline travel agents, from which users seek advice on a possible holiday destination. From a technical viewpoint,

these systems primarily adopt a content-based approach, in which the user expresses their needs, benefits, and constraints using predefined features/attributes. The system then matches the user preferences with travel services in a catalogue of destinations. As such, they are inherently limited in generating recommendations on the basis of the existing user profile; thus, they are ineffective in promoting new destinations or new tourist activities.

Other popular travel sites, such as Expedia.com.sg and TravelAdvisor.com.sg, use web query interfaces to ascertain customer needs before retrieving data from relevant web databases. Typically, a customer wanting to book a hotel would have to search multiple sites to find the best value hotels. Given the increasing number of travel sites, this process is usually tedious, frustrating, and time-consuming. As a better but more challenging alternative, the authors of [2] proposes constructing a global query interface that allows uniform access to disparate relevant sources. The customer indicates their requirements in a single global interface, and all the underlying sources (or databases) are automatically searched and scraped. The retrieved results can then be integrated as shown in Figure 1.
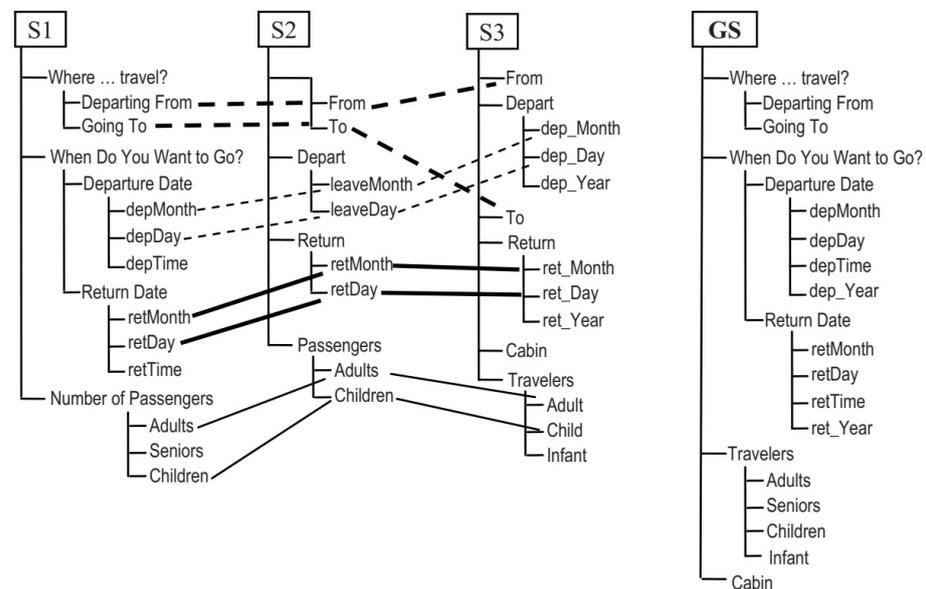
Figure 1. Three input query interfaces (S1, S2, and S3) and global query interface derived (GS) [2].

A good TRS is expected to provide broad suggestions to the users and allow them to choose their travel routes along with activities. The usage of multiple techniques to filter the activities for generating the recommendation is a new scope in this domain. These recommendations are manifested by the TRS functionalities, which may include tour packages, list of attractions, itinerary, and social media capabilities.

- A tour package may include flights, hotel, and accommodations, as well as tourist attractions [39]. TRS of this type are used by travel agencies to find suitable travel packages for customers. Examples include PersonalTour [40], Itchy Feet [41], MyTravelPal [42], and Traveller [43].
- Tourist attractions, temporal events, and other places of interest are usually ranked on the basis of destination, budget, and other information provided by the user [44]. Content-based and contextual analysis are usually considered for the recommendations in these types of TRS such as Turist@ [45]. Examples in the Singapore context include TripAdvisor's Things to Do in Singapore [46], Singapore tourist attractions reviews by local experts—TheSmartLocal [47], and MakeMyTrip's Places to Visit in Singapore [48].
- Planning a travel route guides the user in preparing a route plan to include several places [49]. Some examples are CT Planner [50], City Trip Planner, e-Tourism, and Otium.

- Social functionalities allow users to interact and share information with other tourists [51]. For instance, Itchy Feet and MoreTourism allow users to organize events or activities with similar tourists apart from interacting and commenting.

A comparison of different interfaces (web, mobile, and hybrid) and functionalities (tourist attraction, destination, trip planner, social features, and context-aware) used in TRS are provided in [52].

Recent developments in TRS point to the increasing use of multimedia content and data mining techniques to predict and make recommendations. For instance, by using the sharable content object reference model [53], the TRS collates information related to the recommendation (photos and videos) and converts them into either a Flash movie or a synchronized multimedia integration language presentation.

Another area currently being explored is the utilization of social data and tools to differentiate cluster of users from cluster of items [52,54]. This possibility allows the use of collaborative filtering by using the data collected for a social recommender system in TRS. For example, TripAdvisor connects users via Facebook to their friends and shares relevant content about where their friends have traveled and where they would like to visit in the future.

3.2. Analytics Approaches in TRS

In TRS, the numbers of users and items are very large. As a result, traditional TRS utilizes partial information for identifying similar attributes of users. In recent years, researchers have expressed varying views on how to employ big data and data mining techniques, as well as social network data, to enhance traditional TRS with better prediction and improved accuracy. Some of the common modeling techniques suggested in these studies are the following:

- Filtering techniques: content-based, demographic, and collaborative [44];
- Model-based: genetic algorithm, neural networks, fuzzy logic, Bayesian networks, singular value decomposition, and latent semantic indexing [55–57];
- Hybrid: user-to-user k-nearest neighbors and item-to-item k-nearest neighbors [58];
- Memory-based: similarity measures and aggregation approaches [59].

A summary of techniques used in modeling TRS was provided by [52]. As can be observed, topic-based context-aware travel recommendation systems are gaining traction due to the accessibility of mobile phone and photo-sharing websites with huge volumes of community-contributed geotagged photos [49,60].

Despite the attention and studies reported by other researchers, TRS is still in its early stage of development and has not achieved the level of maturity in other domains such as book purchase and movie recommendations. As such, there remain significant improvement opportunities [61,62], some of which we plan to tackle during the second phase of our project, as outlined in the next section.

4. Proposed TRS Model and Prototype Development

4.1. TRS Modeling Requirements

Leveraging analytics to drive competitive advantage is now a cliché. The usage and the application of big data, along with appropriate analytical techniques, are no longer aimed at driving incremental benefits, but rather inspiring new ways of transforming processes, organizations, or even an entire industry [19]. Deep analysis of a customer's buying patterns allows organizations to move beyond mass marketing and toward more relevant and persona level targeted marketing tactics. Such analytical insights can help develop self-service capabilities to meet individual customer needs and identify new product and service opportunities.

Context-aware recommendation of personalized tourism resources has been made possible by improvements in computing capabilities and the invention of powerful filtering algorithms which can match a user's profile in the form of interests and contexts against a

large knowledge base of tourism resources [7]. However, this knowledge base requires a collaborative development approach due to the heterogeneity, volume, and dynamic nature of the underlying resources as outlined in Section 2. In [63], the authors mined information from social media, community-based photographs, and user experience to come up with the final top k query algorithm. Inspired by these studies and specifically by the framework presented in [16], we first categorized the potential information modeling from big data sources according to their permanence in determining travel recommendations, as shown in Figure 2.
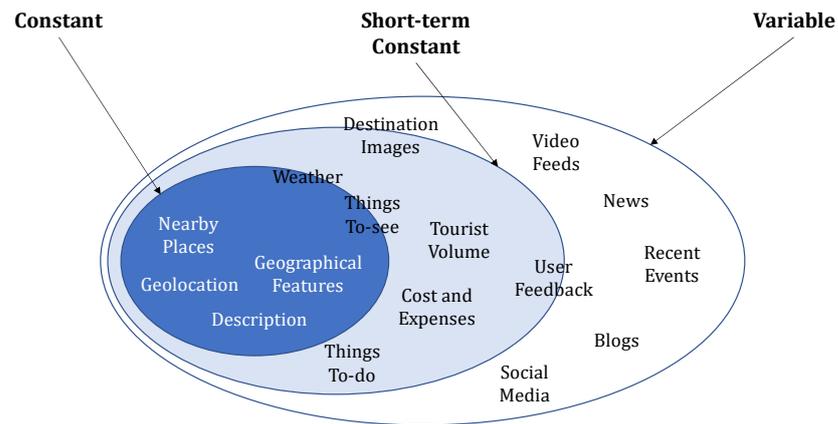
Figure 2. Modeling of information requirements from big data sources for TRS.

It should be noted here that any user-related information, such as demographics, past purchases, or interaction histories were kept outside of the scope for prototype development. This is because such information is generally internal to the companies, and obtaining a sizeable and representative dataset may pose a practical challenge at this stage. The solution, however, was planned in a way that can easily augment such user-related information in the future to drive more refined recommendations. It should also be noted that all these different types of information need not be gathered from different sources. Information such as "nearby places", "things to see", and "things to do" may all be available or derivable directly from the same destination description. Keeping this in perspective, we also identified the potential sources to gather and create all the key information, as outlined in Table 3.

Table 3. Mapping of data sources with big data characteristics and their usage in TRS Model.

Source	Volume	Velocity	Variety	Veracity	Value	Usage in TRS Model
TripAdvisor	High	Medium	Text and Numeric	High	High	User feedback and sentiment mining
Lonely Planet	Medium	Low	Text	High	High	Description, nearby places, geographical features
Twitter	High	High	Text	Medi-um	Medium	Recent events
FlickR	Medium	Medium	Image	High	Medium	Things to see/do
Tutiempo	Medium	Low	Numeric	High	Medium	Weather information
Latlong	Low	-	Numeric	High	High	Geolocation, nearby places
Lonely Planet	Medium	Low	Text	High	High	Description, nearby places, geographical features
Twitter	High	High	Text	Medi-um	Medium	Recent events
FlickR	Medium	Medium	Image	High	Medium	Things to see/do
Tutiempo	Medium	Low	Numeric	High	Medium	Weather information
Latlong	Low	-	Numeric	High	High	Geolocation, nearby places

Once the relevant sources were finalized, after a brief evaluation of the quality and quantity of available information, we considered overlaying the frameworks outlined in [4,15]. We arrived at our proposed TRS model consisting of eight stages, as shown in Figure 3.
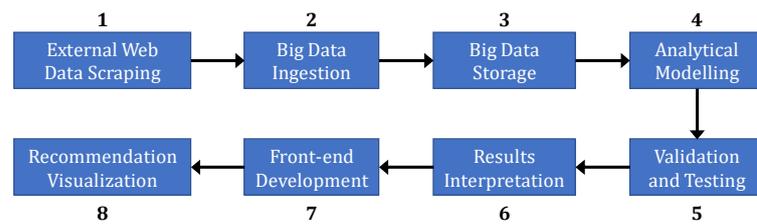
Figure 3. Detailed stages in proposed TRS model.

Our TRS model is proposed with independent modules for different data types from different big data sources up to stage 5. These include mining the information gathered and preprocessing from each source using suitable techniques. The analytical techniques are decided on the basis of the related studies. Given below are some techniques adopted for a different context.

(i)   In [62], the authors used topic modeling on user reviews using latent Dirichlet allocations to create destination attributes or aspect information before performing sentiment mining;
(ii)   In [63], the authors included detailed step-by-step text and sentiment mining processes for generating similar insights, although the scope was restricted to hotel reviews only;
(iii)   In [18], the author proposed a "chatter index" to capture the recent events from Twitter data;
(iv)   In [24], the authors used demographics to complement image-inferred attributes.

While some of the existing studies focused on extracting user attributes to supplement demographic features, we plan to use the same concept for destinations, i.e., to extract destination-related features and augment them with the attributes obtained from other data sources.

Once each module provides satisfactory results, our model uses triangulation and interpretation of these results in stage 6 to create a consolidated list of attributes describing each destination. These attributes include the activities, offerings, geographical features, or other relevant information that are pertinent for a user's decision making. Similar to the findings in [10], we expect these attributes to help create a "destination image" in the user's mind. A key benefit of this modular approach is the flexibility of addition or deletion of big data sources in future providing scope for scaling, without disrupting the functioning of the TRS. The output of stage 6 is a structured data matrix where each row represents a destination and the columns capturing the degree of fulfilment of different attributes for that destination, e.g., for Singapore as a destination, the attribute "beach" may have a value of 1, whereas the attribute "mountain" would have a value of 0. Furthermore, our TRS is modeled to create normalized fulfilment values for each attribute in the [0, 1] range such that these can directly be interpreted as probabilities.

Once this data matrix is created, the recommendation system is ready to interact with the user through a front-end system, as captured in stage 7. The purpose of the front-end system is to input user requirements and preferences, which are then mapped to these predefined attributes. Once mapped, the most relevant destinations satisfying the user needs are looked up from stage 6 output and are displayed to the user. An appropriate visualization scheme [64] for displaying these top destinations matching with user preferences and constraints [65] forms the final stage 8 of our proposed TRS model.

In a nutshell, the stages in our proposed TRS model follow a high-level big data architecture with an end-to-end data pipeline, as depicted in Figure 4. Most of the input data are sourced through web crawling and scraping, which are then preprocessed, ingested, and integrated the data processing (analytics layer) to create a knowledge base for each destination. Once the knowledge base is ready, it is used in an intelligent search-based information retrieval process in the data serving layer, which can interact with the end-user through a presentation component. Figure 4 shows the augmented big data analytical model we employed for TRS.
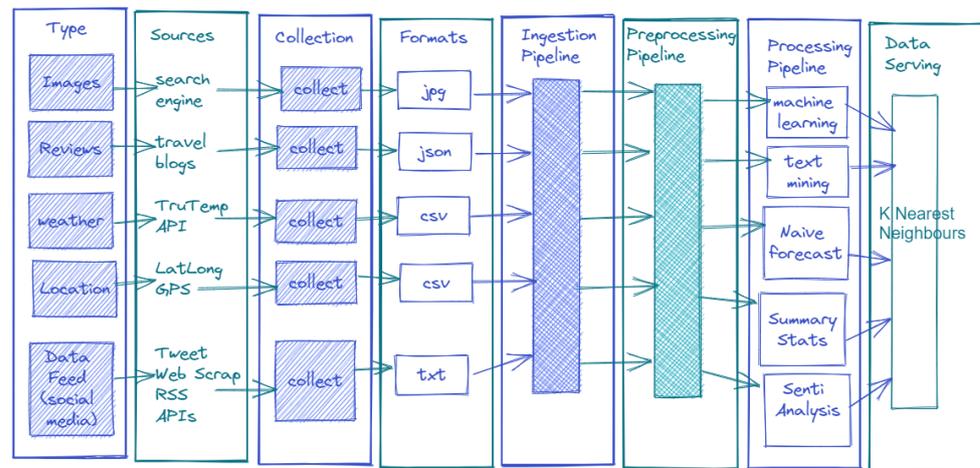
Figure 4. Augmented big data analytical model for TRS.

4.2. Prototype Development

In this section, we describe the prototype development of a thematic TRS using our proposed augmented big data analytical model. The goal is to take into account user preferences, dynamic contexts, required activities, lifestyle experiences, and practical concerns (e.g., cost and distance) to identify and recommend the most suitable set of destinations with a best fit. Such a system would demonstrate a vast improvement in the recommender systems used in the existing commercial systems that focus primarily on tourist attractions offered around packaged destinations that lack meeting the user-centered and context-driven requirements. Furthermore, a single data source is not sufficient to get a holistic rich information about any travel destination.

To overcome the abovementioned drawbacks in existing systems, we develop a prototype of our proposed TRS using the augmented big data analytics model by considering five major categories of data types coming from user-centered and context-driven input sources or themes: (i) images, (ii) reviews, (iii) climate, (iv) social media, and (v) location. We make use of information related to destinations such as images of natural surroundings, reviews on various tourist activities, climate based on history of weather reports, social media content from recent events and global news, and location with geospatial distance measures and user-centric travel constraints. We describe the application of our proposed model in each of these categories of data sources by employing intelligent analytical techniques and state-of-the-art technologies for achieving an enhanced thematic TRS. Figure 5 gives an overview of our augmented content/feature-based recommendation system for the prototype development of our thematic TRS.
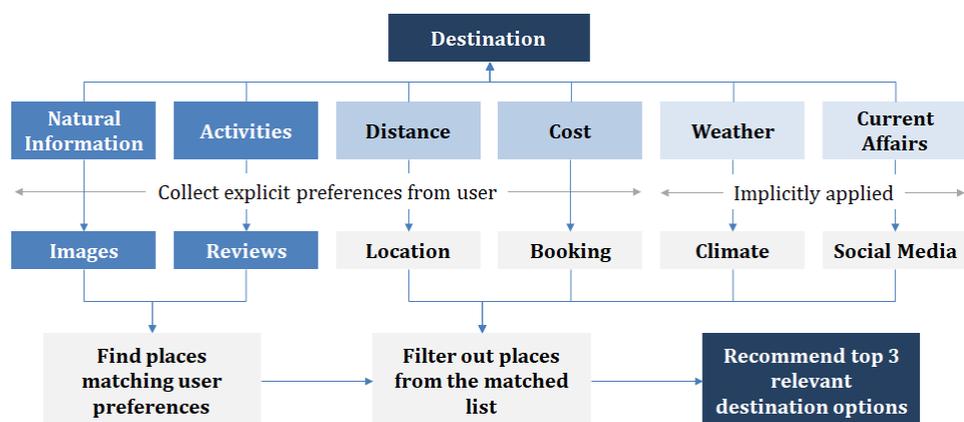
Figure 5. Overview of prototype development of thematic TRS.

We selected 81 destinations covering 12 different Asian countries for the prototype development. For the natural information category, we included images of mountain, forest, beach, city, and village surrounding each destination. For the reviews category, we included activities on wildlife, hiking, cruise, snorkeling, water sports, spa, night life, and family friendly aspects. Three climate indicators (hot and humid, rainy, and cold) were derived from various weather parameters, and sentiment scores were developed for recent events and global news that could have an impact on travel to the destinations.

We provide an overview of the big data analytics process flow adopted for each theme available under the categories of images, reviews, climate, social media, and location for the prototype development of TRS. We describe three key steps (data collection, data ingestion, and data analytics) adopted for each of these five themes for developing the prototype. Lastly, we demonstrate the process of integrating these individual components into an augmented recommendation system for outputting the top three recommended destinations that are aligned closely with user preferences. We employed GitHub and other online public repositories for this purpose. The details of the prototype development are given below.

(i) Big Data Analytics of Images for Thematic TRS

An overview of the main stages involved in processing image data in the augmented analytical modeling of our proposed TRS model is provided in Figure 6. Details of the different steps are given subsequently.

Figure 6. Big data analytics process flow of images for thematic TRS.

(a) Data collection

We employed Python programs to scrape images related to mountain, beach, forest, city, and village from public repositories and portals. These images were preprocessed to extract the features into a training dataset, which was required for deep learning and further processing.

(b) Data ingestion

Using Hadoop, we performed data ingestion of image datasets that were normalized, and operations such as rotation, vertical flip, zoom, or channel shift were employed for data transformation. These included data augmentation with weights for transfer learning from public datasets such as Imagenet using Python and Keras.

(c) Data analytics

The data analytics step was developed using the supercomputing infrastructure of National Supercomputing Center Singapore (NSCC), along with publicly available resources such as Keras, Tensorflow, ResNet50, and Google Colab. For transfer learning of the Imagenet in Keras, all convolutional neural network (CNN) layers were used. A deep learning ResNet CNN model was developed in Python with the transfer learning approach to divide image datasets into five categories *mountain, beach, forest, city, and village), with weights to extract the image features, replacing the top layers into a flattened layer, and a neural network layer for prediction. Our model was trained with about 200 images for each category. The destination images were classified using the trained CNN model to get probabilities of each attribute for each destination.

The model training involved, firstly, freezing all CNN layers to train only the top layers for two epochs. Next, unfreezing all CNN layers was performed to train the whole model using a very small learning rate along with fine-tuning of the weights for 10 epochs. By maintaining the latest model with weights after each epoch, we compare according to the best validation accuracy to arrive at the best model and weights with the highest accuracy. Our final model achieved a high validation accuracy of 96.94 %. Furthermore, prediction of the destination category based on attributes identified in the images was

used to provide insights on the dominant natural features of the destination. For instance, the results achieved from Table 4 show that, for Bali, the order of predictions was beach > village > mountain, while, for Chengdu, the order of prediction was city > village, which was successfully validated.

Table 4. Prediction of destination categories using CNN Model.

Destination	Mountain	Beach	Forest	City	Village
Agra	0	0	0.0625	0.9375	0
Amristar	0	0.10526315789473684	0	0.8421052631578947	0.05263157894736842
Bagan	0.4117647058823529	0.11764705882352941	0	0.4117647058823529	0.0588235294411764705
Bali	0	0.6666666666666666	0	0.06666666666666667	0.26666666666666666
BandaAceh	0.05	0.35	0.05	0.35	0.2
Bandung	0.125	0	0.0625	0.5625	0.25
Battambang	0	0.21052631578947367	0.15789473684210525	0.3684210526315789	0.2631578947368421
Beijing	0	0.2	0	0.8	0
Bhaktapur	0	0	0	1.0	0
Bintan	0.058823529411764705	0.6470588235294118	0	0.058823529411764705	0.23529411764705882
Boracay	0	0.8333333333333334	0.08333333333333333	0	0.08333333333333333
CebuCity	0	0.1	0	0.7	0.2
Chengdu	0	0	0	0.85	0.15

(ii) Big Data Analytics of Reviews for Thematic TRS

An overview of the analytical modeling for processing destination reviews in the thematic TRS prototype development is shown in Figure 7. Details of the different steps are summarized subsequently.

Figure 7. Big data analytics process flow of reviews for thematic TRS.

(a) Data collection

We extracted reviews from web portals such as TripAdvisor for the destinations as JSON files using Python programs as given in Figure 8.

Figure 8. List of JSON files using Python programs to extract reviews from web portals.

An example of reviews extracted for Singapore as the destination is shown in Figure 9.

Figure 9. Example reviews extracted for a destination.

A combination of tools such as Java, Scala, Spark, and Hadoop were adopted, and the environment variables were configured for further data processing.

(b) Data ingestion

In the data ingestion step, a customized data dictionary of keywords (Figure 10) from reviews characterizing each destination attribute was imported as a Spark resilient distributed dataset (RDD).


```
Wildlife, animals, birds, insects,zoo

Hiking, hike, long walk, long walks

Cruise, cruises, cruised, ship, boat

Snorkelling, snorkel,underwater,seaworld

Watersports, Aquatics, rowing, kayak, sailing

Spa, massage

Nightlife, bar, evening market

Familiy friendly, familiy, children, son, daughter
```

Figure 10. Data dictionary of keywords of a destination's characteristics.

(c) Data analytics

A count of the activities was created with two user-defined functions (UDFs) that measured the number of JSON files and the count of keywords extracted from the reviews. A screenshot shown in Figure 11 provides samples of these.

Figure 11. A sample screenshot of functions created and count of keywords extracted.

In the data analytics step, the proportion of each activity in terms of importance and relevance was derived from reviews, and a summary of outputs is given in Figure 12 as an illustration.

	Destination	Wildlife	Hiking	Cruise	Snorkelling	Watersports	Spa	NightLife	FamilyFriendly
1	Destination	Wildlife	Hiking	Cruise	Snorkelling	Watersports	Spa	NightLife	FamilyFriendly
2	Agra								
3	Amritsar								
4	Bagan	0	0	0.214285714	0	0.035714286	0.25	0.107142857	0.392857143
5	Bali	0	0.111111111	0.111111111	0	0	0	0.111111111	0.666666667
6	Banda Aceh	0	0	0.15	0.45	0	0.05	0.05	0.3
7	Bandung	0.020833333	0.0625	0	0	0	0.270833333	0	0.645833333
8	Bangkok	0.047619048	0	0.166666667	0	0	0.19047619	0.142857143	0.452380952
9	Batam	0	0	0.177419355	0.016129032	0	0.193548387	0.080645161	0.532258065
10	Battambang	0.029411765	0.029411765	0.264705882	0	0	0.147058824	0.088235294	0.441176471
11	Beijing	0	0.392857143	0.142857143	0	0	0	0	0.464285714
12	Bhaktapur	0	0.051724138	0	0	0	0.017241379	0.724137931	0.206896552
13	Bintan	0.230769231	0.030769231	0.076923077	0.046153846	0.061538462	0.030769231	0.061538462	0.461538462

Figure 12. An output of thematic analysis of destination reviews.

(iii) Big Data Analytics of Climate Information for Thematic TRS.

An overview of the analytical modeling adopted for processing a travel destination's climate information is shown in Figure 13. Details of the different steps are described subsequently.

Figure 13. Big data analytics process flow of climate information for thematic TRS.

(a) Data collection

We considered publicly available climate information for various destinations in the world with historical data, in some cases dating back to 1929. We included appropriate data sources for climate data websites such as tutiempo.net/that contain historical weather data collected on a daily basis. Various data fields such as temperature, pressure, humidity, precipitation, and wind speed were scraped from the website. Data cleansing and standardization in terms of data format and string symbols were performed for further processing.

(b) Data ingestion

In the data ingestion step, the cleaned climate dataset was uploaded into HDFS using Scala IDE to create a schema of 20 data fields representing various climate information, as Figure 14.

```
|-- continent_name: string (nullable = true)
|-- country_name: string (nullable = true)
|-- city_name: string (nullable = true)
|-- year_name: integer (nullable = true)
|-- month_name: string (nullable = true)
|-- day_name: integer (nullable = true)
|-- T: string (nullable = true)
|-- TMax: string (nullable = true)
|-- Tmin: string (nullable = true)
|-- SLP: string (nullable = true)
|-- H: string (nullable = true)
|-- PP: string (nullable = true)
|-- VV: string (nullable = true)
|-- V: string (nullable = true)
|-- VM: string (nullable = true)
|-- VG: string (nullable = true)
|-- RA: string (nullable = true)
|-- SN: string (nullable = true)
|-- TS: string (nullable = true)
|-- FG: string (nullable = true)
```

Figure 14. An example schema representing climate information.

(c) Data analytics

Statistical software tools were used to address missing data merge and to recalibrate the data range for arriving at an overview of the climate data and distribution for each destination to enable data analysis on a monthly basis. Figure 15 provides a sample output showing a normalized summary of climate data of a travel destination for each month.

```
|   city_name|April|August|December|February|January|July|June|March|May|November|October|September|
+------------+-----+------+--------+--------+-------+----+----+-----+---+--------+-------+---------+
|  Kutchan-cho|  300|   310|     310|     283|    341| 310| 300|  310|310|     300|    310|      300|
|      Pokhara|  120|   155|     186|     113|    186| 124| 120|  124|124|     150|    155|      150|
|        Kyoto|  300|   310|     310|     283|    341| 310| 300|  310|310|     300|    310|      300|
|    Singapore|   90|    93|      93|      85|     93|  93|  90|   93| 93|      90|     93|       90|
|    Jaisalmer|  300|   310|     310|     283|    341| 310| 300|  310|310|     300|    310|      300|
|   Hatsukaichi|  300|   310|     310|     283|    341| 310| 300|  310|310|     300|    310|      300|
| Kuala Lumpur|  300|   310|     310|     283|    341| 310| 300|  310|310|     300|    310|      300|
|       Mumbai|  300|   310|     310|     283|    341| 310| 300|  310|310|     300|    310|      300|
|          Goa|  300|   310|     310|     283|    341| 310| 300|  310|310|     300|    310|      300|
| Penang Island|  240|   217|     217|     226|    248| 217| 240|  248|248|     210|    217|      210|
|         Nara|  300|   310|     310|     283|    341| 310| 300|  310|310|     300|    310|      300|
|      Lumbini|  120|   155|     155|     113|    186| 155| 120|  124|124|     150|    155|      150|
|        Pakse|  300|   310|     310|     283|    341| 310| 300|  310|310|     300|    310|      300|
|     Langkawi|  300|   310|     310|     283|    341| 310| 300|  310|310|     300|    310|      300|
| Luang Prabang|  300|   310|     310|     283|    341| 310| 300|  310|310|     300|    310|      300|
|      Sauraha|  120|   155|     186|     113|    186| 124| 120|  124|124|     150|    155|      150|
|     Amritsar|  300|   310|     310|     283|    341| 310| 300|  310|310|     300|    310|      300|
|    Kathmandu|  300|   310|     310|     283|    341| 310| 300|  310|310|     300|    310|      300|
|    Hong Kong|   90|    93|      93|      85|     93|  93|  90|   93| 93|      90|     93|       90|
|      Boracay|  300|   310|     310|     283|    341| 310| 300|  310|310|     300|    310|      300|
+------------+-----+------+--------+--------+-------+----+----+-----+---+--------+-------+---------+
```

Figure 15. A summarized output showing climate data of a travel destination.

For the prototype development, we focused on three climate indicators (temperature, humidity and precipitation) and applied time series modeling using packages such as the Cloudera Spark Time Series package and naïve forecast techniques for weather forecast. Using Apache Zeppelin, the data analytics of monthly temperature of an example destination, Kutchan-cho, could be visualized, as shown in Figure 16.
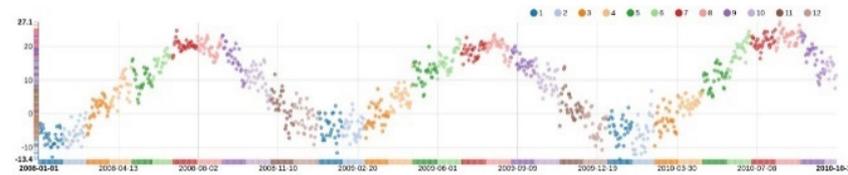
Figure 16. Visualization of monthly temperature in Kutchan-cho.

(iv) Big Data Analytics of Social Media Information for Thematic TRS

An overview of the analytical modeling used in our thematic TRS prototype development for processing social media information from websites such as Twitter is shown in Figure 17. Details of the different steps are captured subsequently.

Figure 17. Big data analytics of social media information for thematic TRS.

(a) Data collection

We employed Apache Flume and Rapid Miner software tools to generate Twitter data files. Rapid Miner Studio was used to create a process to manage connections with Twitter messages or Tweets.

(b) Data ingestion

We employed Apache Flume for data ingestion of tweets into the HDFS path of the Hadoop storage system. Twitter data files were fetched using the consumer key, consumer access token, and access token secret code. Figure 18 shows the data ingestion configuration file.
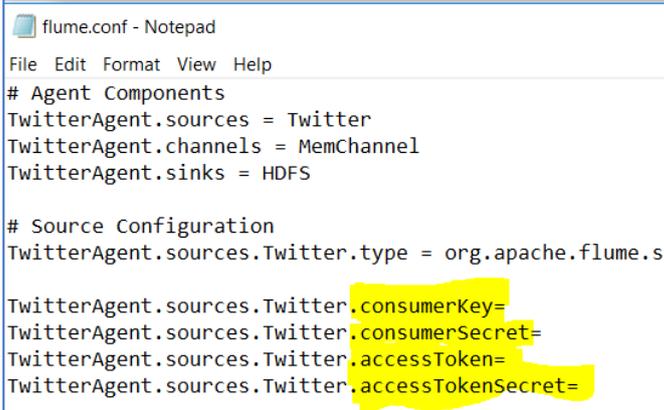
Figure 18. Configuration for data ingestion of Twitter messages.

(c) Data analytics

In the data analytics step, we used the Rapid Miner Radoop extension feature to directly interact with the Hadoop machine as shown in Figure 19. In this way, we avoided the complexity of data preparation and machine learning on Hadoop and Spark. Hive scripts were used to create destination-wise tables of tweets, as shown in Figure 20. We used AFINN, a dictionary consisting of 2500 English words rated from +5 to −5 depending on their meaning, for calculating the sentiments and for performing sentiment analysis of the tweets for each travel destination. For this purpose, user-defined table generating functions (UDTFs) were used to find tweet sentiments and summarize them as a score. The average rating of each tweet was calculated, and the scores were normalized.
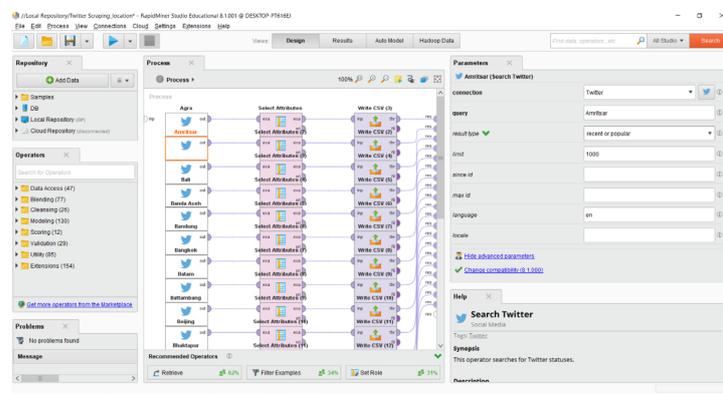
Figure 19. Rapid Miner Radoop for direct analysis using Hadoop storage of tweets.
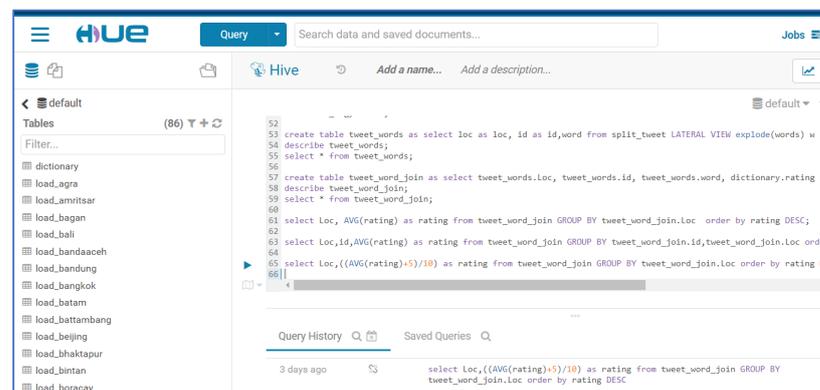
Figure 20. Twitter messages sored in HDFS using HUE with Hive scripts.

(v) Big Data Analytics of Location Information for Thematic TRS

Location information, such as latitude and longitude of a travel destination, is predominantly static. However, we provide the data processing adopted for the prototype development using location-based information for our proposed thematic TRS.

(a) Data collection

The static information about the location of each travel destination was readily available from public sources and was used for calculating pairwise distance measures. However, other dynamic information such as cost per night stay at a location and the travel costs and time of journey between locations were required to meet the traveler's requirements within various constraints.

(b) Data ingestion

Data ingestion of various static and dynamic information related to the location of destinations and traveler requirements and constraints was performed using the Hadoop user interface (HUE). The features of HUE were used to effectively explore and analyze data.

(c) Data analytics

The pairwise distance measure was calculated between each pair of destinations using the following set of formulas:

$$a = \sin^2(\Delta\varphi/2) + \cos\varphi1 \cdot \cos\varphi2 \cdot \sin^2(\Delta\lambda/2), \tag{1}$$

$$c = 2 \cdot \text{atan2}(\sqrt{a}, \sqrt{(1-a)}), \tag{2}$$

$$d = R \cdot (\pi-c), \tag{3}$$

where φ denotes the latitude, λ denotes the longitude, R is the Earth's radius (mean radius = 6371 km), and the angles for the trigonometric functions are in radians. Flight durations could also be calculated assuming an average 800 km/h speed of aircrafts. However, these could be obtained dynamically on the basis of the total time taken by each flight carrier, which could vary in actual instances. The actual flight timings and the availability of direct flights or connecting flights, as well as costs associated with different times of day, were processed. We restricted our model testing to only Asian destinations for the prototype development.

5. Results and Discussion

The outputs generated from the big data analytical models of the five themes of travel data (images, reviews, climate, location, and social media) were integrated in the prototype development of our thematic TRS, as shown in Figure 21.
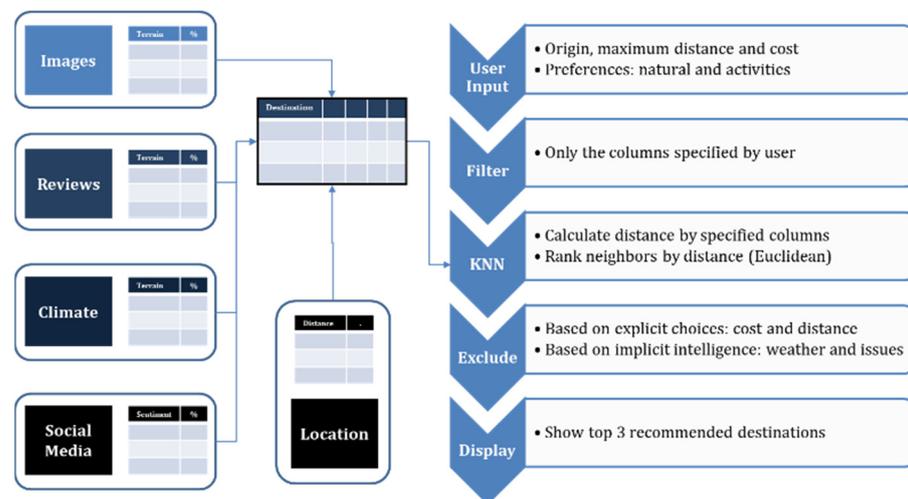

Figure 21. Integration of the augmented big data analytics.

For our prototype development, we successfully implemented the integration of big data analytics of natural information from images with five geographic features, reviews of at least eight activities, monthly weather forecast for hot and humid, cold, and rainy climate, social media-based sentiment summaries, and location features including distance, cost, and description of 81 different travel destinations. For the integration module, Spark and Scala were employed, and quality checks were performed to ensure correctness in features and logical values derived and predicted using each dataset. Figure 22 gives an illustrative summary of the integrated data derived from the five thematic data analytics. The green cells with a value of 1 indicate the positive match of the destination's theme with user's preference, while red cells with a value 0 represent a negative match.

Origin	Month	Max.FlightTime	Max.CostPerDay	Mountain	Beach	Forest	City	Village	Wildlife	Hiking	Cruise	Snorkelling	Watersports	Spa	Nightlife	FamilyFriendly	DummyKey
Singapore	Jun	4	200	0	1	9	9	0	0	0	9	1	1	9	9	9	1
NewDelhi	Dec	5	250	1	0	9	9	9	1	1	0	0	0	9	9	9	1
NewDelhi	Dec	5	250	0	1	9	1	9	9	0	9	9	1	1	9	1	1
Jiuzhaigou	Mar	8	300	9	9	9	9	9	9	9	9	9	9	9	9	9	1
Manila	Aug	5	200	0	0	1	9	9	1	1	0	0	0	9	9	0	1

Figure 22. Summary of an integrated data view of the augmented big data analytics model.

Lastly, Figure 23 shows the prototype's front-end interface for user input and the thematic TRS output displaying the top three recommendation of destinations for the user. Figure 24 shows the output when a user search was made with preferences through the user interface, as shown for beaches within a 4 h flight time from Singapore. The user preferences were for water sports and snorkeling, with no activities related to mountains, village, wildlife, or hiking.
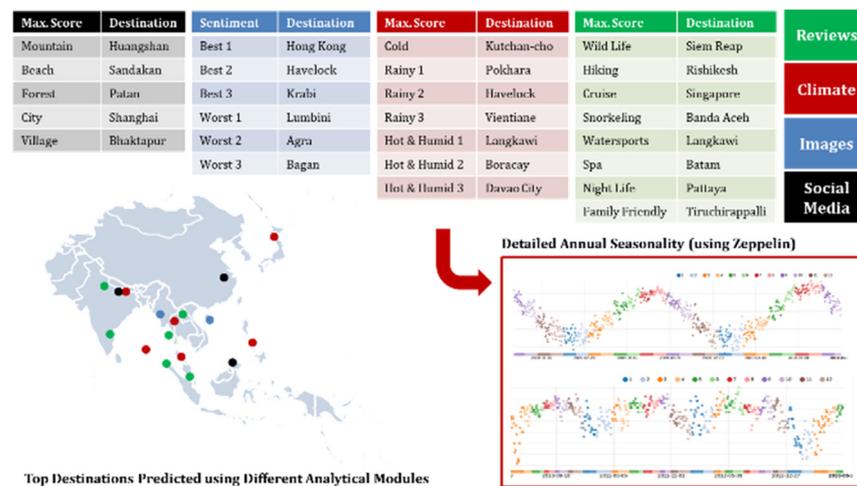

Figure 23. User input and thematic TRS output.

Our prototype development demonstrates a successful application of our proposed thematic TRS using augmented big data analytics as a pilot conceptualization of our model in our ongoing study in this domain. Among the five augmented big data analytical models employed in this study, the image-based prediction models performed well, achieving a high 97% accuracy based on the validation datasets and not specific to the limited destinations considered for the prototype development. On the other hand, custom dictionaries built for processing reviews on each activity require enhancements to achieve more robust and customized results in future. For instance, on the basis of the user preferences and ranking, the recommendations algorithm was able to mine various options. The model used WikiTravel content management system to also crosscheck details such as landmarks and preferences. Two sample suggestions produced are listed in Figures 25 and 26. The

search was performed for beaches within 4-h flight time from Singapore for water sports and snorkeling, along with a user preference for no mountains, village, wildlife, or hiking. The most valued option "Havelock" is a picturesque natural paradise with beautiful white sandy beaches, rich coral reefs, and lush green forest. It is one of the populated islands in the Andaman group. While there are many beaches within 4 h reach, beyond the water sports requirements, Havelock also notably and uniquely satisfies the no preference list. Thus, the proposed recommender system opens new opportunities for matching personalized requirements when retrieving information on the Internet about various popular destinations.
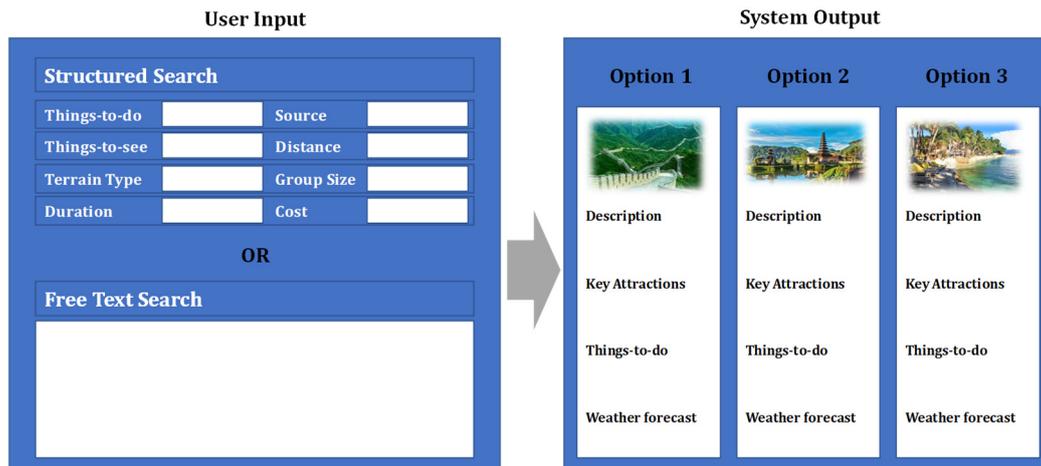

Figure 24. User input and system output of thematic TRS.

Figure 25. Example 1 of destination suggestions from thematic TRS.

Figure 26. Example 2 of destination suggestions from thematic TRS.

In a second example, the search was for family-friendly places within 5 h of flight time from Jiuzhaigou, China, which have a forest and offer a village life experience. Sauraha was a surprise suggestion, as the description of a village is only present in WikiTravel, whereas all other descriptions mention it as a town. However, the algorithm intelligently used text mining, searching, indexing, and weight techniques to present the destination as one of the top suggestions. Thus, the search space of the algorithm is guided by user preference, as well as users' ranking.

In this prototype, we focused on creating an end-to-end pipeline first before expanding the datasets, which would impact the runtime within the cluster significantly. Additionally, categories created on the basis of landscape, natural information (e.g., historic place), and activities helped the recommendation model to provide more options. A cautionary task for the data engineers was to ensure objective and exclusive definition of such categories did not overlap, making the overall recommendation process confusing. This was possible for the selected cleansed dataset, but more munging processes would be required for larger datasets. In addition, reliability of the distributed data storage system could be studied as future work [66]. Optimization of our algorithm can also be considered to recognize the design parameters in an optimal way [67].

5.1. Advantages and Disadvantages of the Model

There are several pros and cons of recommender systems based on the machine learning algorithms of the model adopted. We list below the key advantages and disadvantages of our proposed model for travel recommendation systems through a thorough analysis of our working prototype.

5.1.1. Advantages

- Accuracy: The ratio of user specifications and destination recommendation is typically responsible for the accuracy of travel recommendation system. Our model uses user preferences, ranking of feature choices, and ranking among destinations based on popularity scores from similarity analysis. Such factors directly contribute to better accuracy.
- Efficiency: Efficiency in terms of memory and computational power also depends upon ratio of users and destinations. Hence, if the number of users exceeds the number of destinations, which can happen in most of the tourism recommendation cases, our proposed destination-based recommendations are more reliable in terms of memory and time required to process.
- Stability: Stability of recommendation is related to occurrence and change in the number of users and destinations in the system over time. Although the user population grows over time, the number of destinations remains fairly stable. This helps in focused recommendation with the only expansion made to the feature sets processed by the model.

5.1.2. Disadvantages

- Limited features: The current recommendation model is limited in prototype implementation by the content made available with the features and the type of features of suggested destinations. Domain knowledge is also crucial to make a recommendation. For example, making a destination recommendation requires knowledge beyond images, WikiTravel, and social media review comments.
- Weighted functions: The recommender model uses weighted linear functions for various ranking selection and produces suggestions for destinations by aggregating the output ranks of all destinations. This increases the complexity in the model, burden on the various query parameters, and processing time.

5.2. Nonfunctional Characteristics

Most machine learning recommender models are developed as functional agents. Spark agents can be packaged and distributed throughout the cluster in easy and nonintrusive ways. While functional characteristics of systems are important, we also considered nonfunctional characteristics for prototype development, as they become critical for real-life deployments in order to make the system affordable, easy to use, and accessible. Key nonfunctional characteristics of our prototype and their limitations are listed below.

- Scalability: In the Hadoop CDH platform used, the size of clusters and the volume of data are the influencing parameters for query performance. Typically, adding more

cluster capacity reduces problems due to constraints such as memory limits or disk throughput. On the other hand, larger clusters are more likely to have other kinds of scalability issues, such as a single slow node that causes performance problems for queries. The prototype was limited to testing the recommendation performance in a distributed cluster of one master node.

- Runtime: The Cloudera Runtime 7.1.4 distribution comprises Hadoop, HBase, Hive, Kafka, Knox, Kudu, Oozie, Parquet, Solr, Spark, Sqoop, and Zeppelin.
- Memory: The prototype used 16 GB of RAM and 40 GB of disk space.
- The Cloudera Manager 5.4.0 sensitive data redaction feature addresses the "leakage" of sensitive information into channels unrelated to the flow of data, but not the data stream itself.

The hybrid recommender was built using various Spark models such as naïve Bayes multiclass classifier with an algorithmic complexity of $O(d \times c)$, where d is the feature set, and c is the number of classes in classifier; the trigonometric functions were employed with algorithmic complexity of $O(n)$, where n is the size of data. Deep learning convolutional neural networks (CNNs) and time series were adopted with algorithmic complexity of $O(n)$. Lastly, NLP with sentiment scores using AFINN (sentiment analysis developed by Finn Årup Nielsen) was applied constantly for large datasets. Optimization and in-depth statistical analyses will be considered in future studies in this ongoing research.

6. Conclusions and Future Work

This paper proposed a thematic TRS using an augmented big data analytics model and developed a prototype to demonstrate its practical implementation. It involved the integration of prediction models developed for four main themes, namely, images, reviews, climate, and social media related to destinations, along with the location-based constraints and user preferences. As a pilot study, the TRS implemented with state-of-the-art big data infrastructure provided promising results by suggesting the top three recommendations based on user contexts. In this pilot study, we focused more on creating an end-to-end pipeline for big data analytics before expanding on the datasets for comprehensive testing. We formulated the experimental hypothesis such that the proposed personalized and ranked requirements are to be collected from the user, which are then used along with unstructured travel content and contextual information. We found that our approach improved the recommendation performance compared to other state-of-the-art models in terms of accuracy, alternatives, and precision. The results obtained from experiments on crawled and curated datasets confirmed this.

As part of future work on this topic, an immediate extension will be to use more data to improve each of the prediction models. Additionally, more categories or more items under each category can be created regarding travel and tourism, particularly social media peer review and the user's past behavior/preferences. Even though the prototype does not fully constitute the hypothesized effectiveness, the following steps will be considered in the future to further improve the quality of recommendations:

- Collect more data and test the scalability of the model. Despite implementing the prototype in the Hadoop cluster and using an in-memory framework such as Apache Spark, scalability can be improved by extending the system into a standard platform hosting such as Kubernetes;
- Incorporate more location-sensitive and context-aware information to be processed in the current recommender pipeline;
- Enhance the text mining approach currently implemented by using an explicit semantic analytics model;
- Include more unstructured content beyond WikiTravel such as YouTube backpacker video analytics, TripAdvisor reviews, Wikipedia, and other social media comments or reviews.

References

1. Ricci, F.; Rokach, L.; Shapira, B.; Kantor, P. *Recommender Systems Handbook*; Springer: New York, NY, USA, 2011.
2. Liu, B. *Web Data Mining*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2011.
3. Afsar, M.M.; Crump, T.; Far, B. Reinforcement learning based recommender systems: A survey. *ACM Comput. Surv.* **2022**, *55*, 1–38. [CrossRef]
4. Kontogianni, A.; Alepis, E. Smart tourism: State of the art and literature review for the last six years. *Array* **2020**, *6*, 100020. [CrossRef]
5. Massimo, D.; Ricci, F. Building effective recommender systems for tourists. *AI Mag.* **2022**, *43*, 209–224. [CrossRef]
6. UNWTO. *World Tourism Barometer*; UNWTO: Madrid, Spain, 2022.
7. Leal, F.; Malheiro, B.; Burguillo, J.C. Recommendation of tourism resources supported by crowdsourcing. In *ENTER 2016 PhD Workshop, Proceedings of the International Conference on Information and Communication Technologies in Tourism, Bilbao, Spain, 2–5 February 2016*; Springer: Cham, Switzerland, 2016.
8. Jannach, D.; Pu, P.; Ricci, F.; Zanker, M. Recommender systems: Trends and frontiers. *AI Mag.* **2022**, *43*, 145–150. [CrossRef]
9. Taylan, O.; Alkabaa, A.S.; Yılmaz, M.T. Impact of COVID-19 on G20 countries: Analysis of economic recession using data mining approaches. *Financ. Innov.* **2022**, *8*, 81. [CrossRef]
10. Marine-Roig, E. Measuring destination image through travel reviews in search engines. *Sustainability* **2017**, *9*, 1425. [CrossRef]
11. Esmaeili, L.; Mardani, S.; Golpayegani, S.A.; Madar, Z.Z. A novel tourism recommender system in the context of social commerce. *Expert Syst. Appl.* **2020**, *149*, 113301. [CrossRef]
12. Lim, K.H.; Chan, J.; Karunasekera, S.; Leckie, C. Tour recommendation and trip planning using location-based social media: A survey. *Knowl. Inf. Syst.* **2019**, *60*, 1247–1275. [CrossRef]
13. Roy, D.; Dutta, M. A systematic review and research perspective on recommender systems. *J. Big Data* **2022**, *9*, 59. [CrossRef]
14. Sreejadevi, P.; Swathi, S.R.; Shelbi, J.; Akhil, P.V. A Survey on Various Travel Recommender Systems. *Int. J. Innov. Res. Science Eng. Technol.* **2017**, *6*, 1661–1667.
15. Zheng, X.; Luo, Y.; Xu, Z.; Yu, Q.; Lu, L. Tourism Destination Recommender System for a Cold Start Problem. *KSII Trans. Internet Inf. Syst.* **2016**, *10*, 3192–3212.
16. Chen, M.; Mao, S.; Liu, Y. Big data: A survey. *Mob. Netw. Appl.* **2014**, *19*, 171–209. [CrossRef]
17. De Mauro, A.; Greco, M.; Grimaldi, M. What is big data? A consensual definition and a review of key research topics. In *AIP Conference Proceedings*; American Institute of Physics: College Park, MD, USA, 2015.
18. Vinod, B. Leveraging BIG DATA for competitive advantage in travel. *J. Revenue Pricing Manag.* **2013**, *12*, 96–100. [CrossRef]
19. Wedgwood, K.; Howard, R. Big data and analytics in travel and transportation. In *IBM Big Data and Analytics White Paper*; IBM: New York, NY, USA, 2014.
20. Ayers, R. How Big Data Has Transformed the Travel Agent's Role. Big Data Made Simple 2016. Available online: https://www.crayondata.com/big-data-transformed-travel-agents-role-2// (accessed on 28 February 2018).
21. Felfernig, A.; Gordea, S.; Jannach, D.; Teppan, E. A short survey of recommendation technologies in travel and tourism. *OEGAI J.* **2007**, *25*, 17–22.
22. Niu, T.; Song, M.; Wang, X.; Wang, L. Tourism Destination Recommendation and Marketing Model Analysis Based on Collaborative Filtering Algorithm. *Mob. Inf. Syst.* **2022**, *2022*, 5905490. [CrossRef]
23. OpenJaw. Demystifying Big Data for Travel #4: Prescriptive Analytics. 2018. Available online: https://www.linkedin.com/pulse/prescriptive-analytics-final-frontier-big-data-travel-john-carney (accessed on 28 February 2018).
24. Cheng, A.-J.; Chen, Y.-Y.; Huang, Y.-T.; Hsu, W.H.; Liao, H.-Y.M. Personalized travel recommendation by mining people attributes from community-contributed photos. In Proceedings of the 19th ACM International Conference on Multimedia, Scottsdale, AZ, USA, 28 November 2011–1 December 2011.

25. Song, H.; Liu, H. Predicting tourist demand using big data. In *Analytics in Smart Tourism Design*; Springer: Cham, Switzerland, 2017.

26. Subramaniyaswamy, V.V.; Logesh, V.R.; Indragandhi, V. Intelligent travel recommendation system by mining attributes from community contributed photos. *Procedia Comput. Sci.* **2015**, *50*, 447–455. [CrossRef]

27. Irudeen, R.; Samaraweera, S. Big data solution for Sri Lankan development: A case study from travel and tourism. In Proceedings of the 2013 International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 11–15 December 2013.

28. Vansteenwegen, P.; Souffriau, W.; Berghe, G.V.; Van Oudheusden, D. The City Trip Planner: An expert system for tourists. *Expert Syst. Appl.* **2011**, *38*, 6540–6546. [CrossRef]

29. Sebastia, L.; Garcia, I.; Onaindia, E. e-Tourism: A Tourist Recommendation and Planning Application. In Proceedings of the 20th IEEE International Conference on Tools with Artifical Intelligence, Dayton, OH, USA, 3–5 November 2008.

30. Montejo-Ráez, A.; Perea-Ortega, J.; García-Cumbreras, M.; Martínez-Santiago, F. Otium: A web based planner for tourism and leisure. *Expert Syst. Appl.* **2011**, *38*, 10085–10093. [CrossRef]

31. Simo, E.R.C. Development of a Tourism Recommender System. Master in Artificial Intelligence (UPC-URV-UB). Master's Thesis, The Open University, Milton Keynes, UK, 2012.

32. Hinze, A.; Junmanee, S. Travel Recommendations in a Mobile Tourist Information System. 2005. Available online: https://researchcommons.waikato.ac.nz/bitstream/handle/10289/9413/GI-Proceedings.63-7.pdf?sequence=3 (accessed on 28 February 2018).

33. Averjanova, O.; Ricci, F.; Nguyen, Q.N. Map-based Interaction with a Conversational Mobile Recommender System. In Proceedings of the Second International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, Valencia, Spain, 29 September 2008–4 October 2008.

34. Rey-Lopez, M.; Barragans-Martinez, A.B.; Peleteiro, A. moreTourism: Mobile recommendations for tourism. In Proceedings of the IEEE Intenational Conference on Consumer Electronics, Las Vegas, NV, USA, 9–12 January 2011.

35. Martin, D.; Alzua, A.; Lamsfus, C. A Contextual Geofencing Mobile Tourism Service. In Information and Communication Technologies. 2011. Available online: https://pdfs.semanticscholar.org/cafa/ffe83a9ba1e73156088577ccc7dc6ababa51.pdf (accessed on 10 February 2022).

36. Current Travel Recommender Systems, EuTravel Project, Horizon 2020. 2017. Available online: http://www.eutravelproject.eu/default.aspx?articleID=8100 (accessed on 10 February 2022).

37. TripMatcher, Welcome to TripMatcher. 2018. Available online: https://www.facebook.com/tripmatcher/ (accessed on 10 February 2022).

38. Your Vacation Coach Inc. Your Vacation Coach. 2013. Available online: https://www.facebook.com/Your-Vacation-Coach-226468734200/ (accessed on 10 February 2022).

39. Allaparthi, A. Vacation Package Recommender System. Master's Thesis, Grand Valley State University, Allendale, MI, USA, 2015.

40. Lorenzi, F.; Loh, S.; Abel, M. PersonalTour: A Recommender System for Travel Packages. In Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Lyon, France, 22–27 August 2011.

41. Sharda, N. *Tourism Informatics: Visual Travel Recommender Systems, Social Communities, and User Interface Design: Visual Travel Recommender Systems, Social Communities, and User Interface Design*; IGI Global: Hershey, PA, USA, 2009.

42. Koceski, S.; Petrevska, B. Empirical Evidence of Contribution to E-Tourism by Application of Personalized Tourism Recommendation System. *Sci. Ann. Econ. Bus.* **2012**, *59*, 363–374. [CrossRef]

43. Schiaffino, S.; Amandi, A. Building an expert travel agent as a software agent. *Expert Syst. Appl.* **2009**, *36*, 1291–1299. [CrossRef]

44. Zhi, Y.J.; Wei, G.; Yi, J.S. An Agent Framework of Tourism Recommender System. In *MATEC Web of Conferences*; EDP Sciences: Les Ulis, France, 2016.

45. Batet, M.; Moreno, A.; Sanchez, D.; Isern, D.; Vall, A. Turist@: Agent-based personalised recommendation of tourist activities. *Expert Syst. Appl.* **2012**, *39*, 7319–7329. [CrossRef]

46. TripAdvisor Singapore. Things to Do in Singapore. 2018. Available online: https://www.tripadvisor.com.sg/Attractions-g294265-Activities-Singapore.html (accessed on 1 December 2021).

47. TheSmartLocal Reviews. Available online: https://thesmartlocal.com/?s=reviews (accessed on 10 February 2022).

48. MakeMyTrip.com. Places To Visit in Singapore. 2018. Available online: https://www.makemytrip.com/travel-guide/singapore/places-to-visit.html (accessed on 1 February 2022).

49. Takeshi, K.; Tomoharu, I.; Go, I.; Ko, F. Travel Route Recommendation Using Geotags in Photo Sharing Sites. In Proceedings of the CIKM'10, Toronto, ON, Canada, 26–30 October 2010.

50. Kurata, Y.; Shinagawa, Y.; Hara, T. CT-Planner5: A computer-aided tour planning service which profits both tourists. In Proceedings of the 9th ACM Conference on Recommender Systems (RecSys 2015), Vienna, Austria, 16–20 September 2015.

51. Xue, A.Y. Recommendation Systems for Travel Destination and Departure Time. Ph.D. Thesis, University of Melbourne, Parkville, VIC, Australia, 2015.

52. Logesh, R.; Subramaniyaswamy, V. A Collaborative Location Based Travel Recommendation System through Enhanced Rating Prediction for the Group of Users. *Comput. Intell. Neurosci.* **2016**, *2016*, 1291358.

53. Learning Technologies Group Plc Ltg, SCORM Explained. 2018. Available online: https://scorm.com/scorm-explained/ (accessed on 10 February 2022).

54. Li, T.; Kou, G.; Peng, Y.; Yu, P.S. An Integrated Cluster Detection, Optimization, and Interpretation Approach for Financial Data. *IEEE Trans. Cybern.* **2022**, *52*, 13848–13861. [CrossRef] [PubMed]

55. Cai, G.; Lee, K.; Lee, I. Itinerary recommender system with semantic trajectory pattern mining from geo-tagged photos. *Expert Syst. Appl.* **2018**, *94*, 32–40. [CrossRef]

56. Mahmood, T.; Ricci, F. Learning and Adaptivity in Interactive Recommender Systems. In Proceedings of the ICEC'07, Minneapolis, MN, USA, 19–22 August 2007.

57. Loh, S.; Lorenzi, F.; Saldana, R.; Licthnow, D. A Tourism Recommender System Based on Collaboration and Text Analysis. *Inf. Technol. Tour.* **2004**, *6*, 157–165. [CrossRef]

58. Mehrbakhsh, N.; Karamollah, B.; Mohsend, R.; Vahid, R. A recommender system for tourism industry using cluster ensemble and prediction machine learning techniques. *Comput. Ind. Eng.* **2017**, *109*, 357–368.

59. Binucci, C.; De Luca, F.; Di Giacomo, E.; Liotta, G.; Montecchiani, F. Designing the Content Analyzer of a Travel Recommender System. *Expert Syst. Appl.* **2017**, *87*, 199–208. [CrossRef]

60. Xu, Z.; Chen, L.; Chen, G. Topic based context-aware travel recommendation method exploiting geotagged photos. *Neurocomputing* **2015**, *155*, 99–107. [CrossRef]

61. Reshma, J. Literature Survey on Tourism Recommendation. *Int. J. Innov. Res. Comput. Commun. Eng.* **2017**, *5*, 1916–1922.

62. Santhibaskaran, D.; Bhuvaneswari, L. Customized Travel Recommendation System using Big Data Analytics. *Int. Innov. Res. J. Eng. Technol.* **2017**, *2*, 88–92.

63. Potdar, S.; More, N.S. Efficient Recommendation System based on Keyword for Smart-City with Sentiment Analysis using Hadoop. *IJETT* **2017**, *48*, 17–23. [CrossRef]

64. Ponnada, M.; Sharda, N. A high level model for developing intelligent visual travel recommender systems. *Inf. Commun. Technol. Tour.* **2007**, *2007*, 33–42.

65. Dhelim, S.; Aung, N.; Huang, R.; Ma, J. Personality-aware product recommendation system based on user interests mining and metapath discovery. *IEEE Trans. Comput. Soc. Syst.* **2021**, *8*, 86–98. [CrossRef]

66. Kou, G.; Yi, K.; Xiao, H.; Peng, R. Reliability of a Distributed Data Storage System Considering the External Impacts. In *IEEE Transactions on Reliability*; IEEE: Piscataway, NJ, USA, 2022. [CrossRef]

67. El-kenawy, E.; Abutarboush, H.; Mohamed, A.; Ibrahim, A. Advance Artificial Intelligence Technique for Designing Double T-Shaped Monopole Antenna. *CMC-Comput. Mater. Contin.* **2021**, *69*, 2983–2995. [CrossRef]