ISSN 2395-8618

Recommender System for Tourist Itineraries Based on Aspects Extraction from Reviews Corpora

Liliya Volkova, Elena Yagunova, Ekaterina Pronoza, Alexandra Maslennikova, Danil Bliznuk, Margarita Tokareva, and Ali Abdullaev

***Abstract*—In this paper a recommender system is described which takes a set of venue categories of user's interest into account to form a tourist itinerary throughout a city. The system is focused on user preferences in venue aspects. Techniques of such aspects extraction are developed in this paper, in particular from reviews corpora. User preferences are used to weigh aspects associated with particular sights and restaurants. These filtered venues along with time restrictions are subject to submit into the recommender system. A lightweight ontology is discussed which describes the domains of restaurants and sightseeing knowledge and allows venues comparative analysis to enhance the search for relevant venues. The system designed performs automated planning of tourist itineraries, flexible sights searching, and analysis of venues aspects extracted from reviews in Russian.**

***Index Terms*—Information extraction, lightweight ontology, natural language processing, recommender systems.**

I. Introduction

SUBJECT area of this research is a recommender system for tourist itineraries planning. Provided with venue reviews corpora, the analyzer component extracts aspects defined for museums and restaurants. A lightweight ontology is described, which serves as a semantic resource for estimating venues for a narrow search and further thematic planning. With support of the lightweight ontology, the recommender system forms a route over a selected set of venue categories (e.g., visiting two museums, then a restaurant, then another museum) basing on user preferences for previously extracted aspects. A flexible recommender engine is designed which generates relevant itineraries throughout the city, each accompanied with a map-based route.

The rating approach is the one avoided in this research: generalized ratings are widespread, but apparently not always exact they are. The task of verifying and attributing ratings itself requires a separate study. Moreover, user preferences differ, which is not reflected by ratings on the whole. Therefore, the thematic recommender system is in the focus of this article, including techniques of venue analogs selection.

One of the most popular web search queries is for tourism and trip planning; hence an automated trip planner is in need of. Most existing resources have a number of limitations. According to a survey undertaken, they contain static information only [36], either no thematic routes planning [38], or fixed routes [18]. The most complete solution provides routes flexible planning [37]. No solution including restaurants into agenda was discovered, nor analogues selection in case of absence of venues, exactly matching the query (as stated above, the rating approach is not under consideration).

The system of interest shares with the above-mentioned sites the goal of providing a tourist trip planner. In this research, the recommender system is designed which is focused on users preferences consideration. For flexibility and thematic search of restaurants and sights, the recommender system under design includes the following subsystems, which will be discussed below.

1) Reviews analyzer with aspects extraction for sights and restaurants.
2) Knowledge base for venues (with a lightweight ontology-supported schema [15]).
3) Recommender system:
   a) content-based recommender strategies;
   b) flexible parameterization with user filters;
   c) lightweight-ontology-driven heuristics (apart from route-forming heuristics).
4) Itinerary building, conjugated with route planning and maps API.

Manuscript received on September 2, 2016, accepted on December 7, 2017, published on June 30, 2018.

Liliya Volkova is with the Moscow Institute for Electronics and Mathematics, National Research University Higher School of Economics, Moscow, 101000, Myasnitskaya ul., 20, Russia, and with the Bauman Moscow State Technical University, Moscow, 105005, 2-ya Baumanskaya ul., d. 5, str. 1, Russia (e-mail: liliya@bmstu.ru).

Elena Yagunova, Ekaterina Pronoza, Alexandra Maslennikova, and Danil Bliznuk are with the Saint Petersburg State University, Saint Petersburg, 199034, Universitetskaya nab., d. 7-9, Russia (e-mail: {iagounova.elena, katpronoza, msasha1996}@gmail.com, blizda@outlook.com).

Margarita Tokareva is with the Moscow Institute for Electronics and Mathematics, National Research University Higher School of Economics, Moscow, 101000, Myasnitskaya ul., 20, Russia (e-mail: rit1336@yandex.ru).

Ali Abdullaev is with the Bauman Moscow State Technical University, Moscow, 105005, 2-ya Baumanskaya ul., d. 5, str. 1, Russia (e-mail: klim_sychev@mail.ru).

ISSN 2395-8618

Aspects extraction techniques are developed for further venues automatic estimation, the detailed description is given in chapter II. The first two subsystems require a specific knowledge organization, which is a lightweight ontology [39], see chapter III. The latter two subsystems are described in chapter IV.

The recommender system based on user preferences implies a technique of evaluating venues aspects in terms of natural language. For ex., a sample user likes art, but not modern art, and his tastes are limited to authentic Italian cuisine. The easiest solution can be found when all of the venues of specified kinds are present in the vicinity. The question is what strategy should be built into the recommender system to search for similar venues in case of absence of the exact match. If there is no authentic Italian cuisine, some substitution should be mined with similarity heuristics (be it French cuisine or a café with pizza). In this research, ontological reasoning is considered to be the solution of this problem. Two sets of aspects are defined for sights and restaurants respectfully (see chapter 2), and the recommender system comprises rules deriving from lightweight ontology relations over the aspects mentioned.

II. Aspect-based Restaurant and Museum Information Extraction

A method for Russian reviews corpora analysis (as part of information extraction (IE)) is discussed, which gathers and structures restaurants and museums parameters from users' reviews, and feeds the recommendation system with the data collected. The focus of this chapter is on extracting aspects (so to be referred to).

IE methods, as well as NLP methods in general, are classified into rule-based, statistical and hybrid. The first approach implies using templates and semantic resources (e.g. WordNet-Affect, SentiWordNet, SenticNet), while statistical methods allow solving the task without such resources [27]. For recommender systems, in particular for museums and conterminal fields, three approaches are mostly combined: (1) content-based, (2) aspect-based, and (3) user-based [17], [21], [30], [32]. The only considered traits of the latter approach in this work are the review language and the informant's homeland. Content-based approach involves full consideration of official museums data from different resources; the aspect-based approach comprises analysis of aspects retrieved by automatic and semi-automatic reviews processing. The goal of the IE task in general is to retrieve most aspects extractable within the two approaches, while the focus of this work is on aspects extraction from reviews corpora, in particular on research for key aspects and analysis of their realization types.

The approach towards corpora analysis presented in this paper is based on non-contiguous bigrams and part of speech (POS) distribution analysis [28]. Trigger words dictionaries are obtained by means of the bootstrapping method. The venues can be described with a set of characteristics, for instance, service quality, food quality, cuisine type, price level, noise level, etc. The key aspects are selected below. All of the aspects to be extracted from the reviews are experts-predefined. No techniques of automatic aspects identification were employed, for these would inevitably introduce noise into the IE model. Most examples are dedicated to restaurants IE.

It should be stated that our corpora consist of Russian colloquial texts, and Russian is known for its rich morphology and free word order which complicate its automatic processing. Another complicating factor is that the practice of data adjusting to common recommender systems standards is not yet widespread in Russia, and therefore users' reviews are often not what one would expect them to be (e.g., free narratives are quite common, with no point of reviewing, as opposite to expected). However, according to the results, an information extraction system for Russian can still be successful, especially when based on the ideas obtained from corpora analysis.

A. Restaurant Information Extraction

The hypothesis is that the most important characteristics of a restaurant are service and food quality along with cuisine type, so the analysis is so far focused on these three (and on the extraction of their aspects). This assumption is proved by the distribution of the aspects in the data. These main aspects are discussed in this section, though more aspects can be aggregated within further research for fine-grained detail.

The next assumption is that the proposed IE system can be highly effective despite the difficulties imposed by the structure of a typical Russian restaurant review. The fact is that, when such a review is concerned, the key information about restaurant characteristics does not always lie on the surface. However, tuning models with respect to the results gained during corpus analysis can increase IE system performance.

The corpus analyzed consists of 32525 users' reviews (colloquial texts) about restaurants (4.2 millions of words). The reviews are provided by tulp.ru and dated 2013. A part of the corpus is annotated in a semi-supervised way (first, automatically using a simple keywords-based algorithm, and then manually corrected by two experts). It includes 1025 reviews about 206 restaurants located in the centre of Saint-Petersburg. The list of aspects is given in Table I (the most important aspects related to food quality, cuisine type and service quality frames, are given in bold).

TABLE I
Restaurant Aspects (Examples)

Restaurant Aspects			
Cuisine type	**Service Quality**	Company	Children menu
Food quality	**Staff politeness**	Audience	Kids area
Noise level	**Staff amiability**	Average cheque	Bar
Service speed	Cosiness	Price level	Parking place

ISSN 2395-8618

The task is actually a classification problem, but the classes differ from aspect to aspect. For example, for kids' area and bar aspects there are 2 classes: available and unavailable; and for the aspects related to food and service quality (service speed, food quality, etc.) we define 5 sentiment classes: -2, -1, 0, 1, 2. For each aspect the system should either label a review with one of the possible classes or reject it as irrelevant with respect to the given aspect. As most restaurants characteristics are never mentioned in the reviews, an empirical threshold frequency value of 10% is defined in this research, and aspects mentioned in at least 10% of reviews are considered. Classifiers were only trained for the frequent aspects (they are divided into groups in Table II).

TABLE II
FREQUENT RESTAURANT ASPECTS DISTRIBUTION IN THE CORPUS

Occurrence Percentage	List of Aspects
[85%; 100%]	Food quality (86%)
[55%; 85%)	Service quality (55%)
[25%; 55%)	Staff politeness and amiability, service speed, price level, cosiness
[10%; 25%]	Noise level, crampedness, romantic atmosphere, company

The information extraction task related to food and service quality can be reformulated as sentiment analysis with respect to the restaurant aspects of interest. For the aspects chosen as the most frequent ones, the following classifiers were considered: Naive Bayes (NB), Logistic Regression (LogReg), and Support Vector Machines (SVM) as implemented in scikit-learn [31]. In this paper an illustration of machine learning is given with respect to food and service quality criteria. Since the cuisine type aspect suggests a multilabeling task, in this section machine learning models are only considered with respect to food and service quality.

Since the annotated corpus includes a large amount of missing values, the classification task is divided into two parts: first, a classifier is trained to tell between missing and present values, and then, if the value is present, the classifier is to predict its class. The latter is discussed in detail in this section.

Our baseline feature set consists of unigrams and bigrams (on the lemma-level, only contiguous ones). Trigrams were also considered, but since they did not improve performance much while increasing feature space dimensions, trigrams were excluded from the feature set. The experiments were conducted with two extended features sets. First, only non-contiguous bigrams were added (with window size equal to 3 as it appeared to perform best). In the second set, emoticons and exclamations, predicative-attributive words and key words and expressions were added instead.

To evaluate the models, shuffle 10-fold cross-validation was conducted. Average weighted F1 scores for food and service quality are given in Table III. The weights are calculated as relative frequencies of the classes in the annotated subcorpus.

TABLE III
FOOD AND SERVICE QUALITY F1 SCORES
(BEST AVERAGE WEIGHTED F1 SCORE GIVEN IN BOLD)

Restaurant aspects	Model	Baseline, %	Extended (1), %	Extended (2), %
Food quality	NB	**69.45**	**70.08**	**70.26**
	LogReg	64.24	68.77	68.64
	SVM	63.99	65.57	66.21
Service quality	NB	**64.37**	**68.77**	**65.33**
	LogReg	56.14	65.05	57.90
	SVM	54.30	63.80	56.27

NB appears to be the best among the three classifiers for both aspects, but its basic and extended versions show similar scores while SVM and LogReg extended versions show improvement compared to corresponding basic versions.

In further phases of research other restaurant aspects were also considered (apart from food and service quality and cuisine type described in this paper), and experiments were conducted with different classifiers, such as Multinomial NB, Decision Trees, Random Forests, and Perceptron-based. Optimal combinations of feature and classifier were selected for each frequent aspect [26].

Basing on the experimental data, the suggestion is to recommend LogReg for the classification of informal unstructured Russian texts into those which contain information or opinion about the specific aspect and those which do not.

At sentiment classification task, NB is best for all the aspects. It can be explained by both the nature of the classifier and the data: NB, having high bias, usually behaves better on the small amount of training data, and for food and service quality aspects there are 5 classes of sentiment which makes the amount of training data inside each of the classes rather small. Therefore it might be suggested that NB is good at classifying sentiment in the informal texts on the small training set.

It should be also stated that including emoticons and exclamations into the feature set is not a good idea unless the aspect is service quality. For the other aspects it does not improve F1 or even impairs it [28].

For the service frame, dictionaries do improve the results. But food quality, one of the most important aspects, is best extracted using non-contiguous bigrams which cover a wide variety of the expressions of opinion. Thus, a more elaborate lexicon and dictionaries construction could be one of the promising work areas.

A thorough corpus analysis was conducted based on non-contiguous bigrams and POS-distribution of the trigger words context. Experiments with several classifiers showed that their performance can be improved with the results and ideas de-

ISSN 2395-8618

rived from corpus analysis, thus proving the importance of the latter. In particular, it has been shown that using trigger words and predicative-attributive words dictionaries is an effective approach for food quality extraction while service quality aspect, which is harder to deal with, demands a wider range of features.

B. Museums Information Extraction

As the recommender system at its origin is dedicated to cultural journeys, the museum topic requires corresponding aspects extraction as well. The implementation of an aspects extraction module necessitates reviews corpus analysis, patterns construction (including development of the methodology for such construction) and evaluation. The approach for patterns construction presented in this paper is based on n-grams (n ranges from 1 to 8) and POS-distribution analysis. Trigger words dictionary and predicative-attributive dictionaries are obtained by means of the bootstrapping method, targeted at the aspects of interest [27], [28].

The key distinctions for museum IE are the vast repertoire for aspects and the main focus on estimating trigger words and patterns coverage of users' reviews. This leads to combining information extraction, opinion mining and sentiment analysis procedures. The implemented approach is based on foresaid results for restaurant IE. But in this paper there is no results discussion for the evaluation stage is ongoing.

At this point the system is based on the following reviews corpus: The State Hermitage – 2 100 reviews, The Museo del Prado – 1 000, The Louvre Museum – 1 525, The Uffizy Gallery – 450, The Rijksmuseum – 425, The National Gallery – 350.

The approach being as for restaurants, the procedure of analysis comprises the following stages: (1) corpus preprocessing (tokenization, lemmatization, normalization, splitting into sentences, filling frequency and n-grams dictionaries), (2) filling nominations and predicative-attributive dictionaries, (3) filling keywords and keyphrases dictionaries, (4) filling modifiers dictionaries, (5) titles analysis for generalized description.

The predicative-attributive dictionaries were chosen, in particular for adjectives and full and short participles, which refer to nominations of the key frames. This is conditioned by the POS distribution analysis within corpus n-grams showed dominating of noun phrases in most aspects description [28].

Aspects are objective (for ex., tickets e-booking, student prices), subjective (for ex., queues for tickets, crowds inside museums), and mixed. The first category requires IE, the second – opinion mining and sentiment analysis, while the last category requires the composition of both approaches. The latter triade is solved in this research: subjective aspects are of interest, these are represented in 5-degree scale (ranged from -2 to 2), namely ticket prices, queues and crowds in museums.

The important point is to make sure that enough cases for aspects under consideration are present in the subcorpus dedicated to one aspect. The threshold is empirical and is 10 %; it would be also actual for the next stage of pre-processing based on machine learning with different classifiers. To compensate weak accessibility of semantic resources for Russian, semi-automatic dictionaries filling is used basing on the reviews corpus, while thorough syntactic analysis is substituted by n-grams analysis (n ranges from 1 to 8). The latter is supplemented by POS tags, in particular, by POS-filters applied to n-grams components [24]. Different types of negation for Russian might also be covered by the same n-grams.

The data on the above mentioned aspects is present for all of the museums named. Basic museums information is extracted, as well as masterpieces (by name and author) and different services. The worst results are obtained for mining exhibitions, even the long-term ones. Reviews in English have such advantage as their uniform structure compared to reviews in Russian. For the latter the problem is in their rather essay character, for example, these sometimes contain compulsive comparisons to the homeland museums (Hermitage, Tretiakov Gallery, Russian Museum, etc.). Reviews in several corpora are non-uniform and vague, as it is stated in [24]. Using semantic dictionaries and hierarchies thesauri [20], which were semi-automatically or manually filled from reviews corpus, allows improving the quality of most aspects extraction.

The repertoire for topics and aspects is vast: general information, masterpieces, exhibitions, service, tickets prices, e-booking for museums tickets, tickets queues, payment by credit card, opening time, etc. All of these aspects imply thorough IE techniques, and the repertoire allows different kinds of routes: from trips for students with low budget to wealthy tourists, from family tourism with children to big youngsters companies. Considering all of the aspects in the recommender system allows covering a wide range of tourist types, so that the system in production would gain success for its detailed search (with blocking or non-clocking aspects, e.g., no 18+ bars, or preferably parks and family leisure). The aspects extracted are provided with interrelations, which form the lightweight ontology, the latter serving not only as dictionary for aspects extraction, but also for estimating objects within venue categories for thematic itineraries recommending described further.

III. The Lightweight Ontology

In order to describe semantics lying behind data, ontologies can be used in an information integration task to make the content explicit [40]. Addressed to the bottleneck of combining domain experts with ontology engineers in order to build a full-sized ontology, a lightweight ontology is intended to meet the expectations of people who argue in favor of

ISSN 2395-8618

powerful, knowledge-intensive applications based on ontological reasoning [7]. It is presumed that lightweight ontologies are limited in their expressiveness and are mostly focused on a hierarchy of concepts [22], but still they have proven useful, this resonates with the so-called Hendler hypothesis [16]: "A little semantics goes a long way." Besides, the problem of unsupervised ontology learning is still unsolved [23] and is most crucial for languages which still do not have semantic resources thorough enough, e.g. for Russian (though several projects exist [20]).

On the base of analysis conducted, it should be stated that applying and developing a taxonomic model and further a lightweight ontology is a perspective approach towards determining venues similarity (through similarity relation [33]) and solving the problems derived from data insufficiency and incompleteness. A pictorial example of a lightweight ontology employment is as follows: if there is no direct "whisky bar" category match available around user's current location, the system should use rule-based analysis and advise alternatives, e.g. a restaurant with an excellent selection of whisky. The approach allows (1) searching within a database with further application of lightweight-ontology-driven rules of venues extraction in case of absence of match, and (2) pre-mining the data to provide more substitutes (with fuzzy estimation). Additionally, the information on a venue might be processed to match the description of venues satisfactory to the query, but not reachable, e.g., in a given time period [10].

Two domains are covered for further referring to their concepts and interrelations as in corresponding domain of knowledge with agreed meanings and properties [14]: restaurants and sightseeing. Besides, intersections of domain vocabularies can slightly disfigure the results [2]. "Good ontology design, especially for larger projects, does require a degree of modularity. An architecture of multiple ontologies often work together to isolate different work tasks so as to aid better ontology management. Ontology architecture and modularization is a separate topic in its own right" [3], [4].

Though several approaches exist towards automatic converting of classifications into lightweight ontologies [11], still initial expert estimation is of big value and is chosen as the path for this research. Three data sources were considered.

(1) The Foursquare [9] classification which is quite fulfilled but does not contain relations nor all of the parameters necessary (e.g., there is no strict cuisine types classification, and one can find a bakery and restaurants with different types of Chinese cuisine on the same level of abstracts). Such hierarchy requires thorough correctives.

(2) An experts-composed taxonomic model which comprises a thorough classification (designated for this research) and is on the relations adjustment stage.

(3) The set of aspects extracted for museums and restaurants for further lightweight ontology filling (extracted with the above discussed techniques).

The advantages of all of the three items are considered for creating a hybrid model. The characteristics from the latter set of aspects are necessary to complete the first two items, refined and modified. The diverse relations are necessary for various tasks requirements: vertical and horizontal, different types of them (for ex., the "differ" relation for classes might reflect music genre, target audience age, average bill). This allows a more detailed and flexible search oriented on refining the output according to user's query parameters. At this stage the lightweight ontology is as stated in Table IV. The OWL [25] is used, the lightweight ontology is under further development.

TABLE IV
CURRENT LIGHTWEIGHT ONTOLOGY VOLUME

Category	Restaurants domain	Sightseeing domain
Classes and subclasses	32	22
Instances	55	25
Object relations	12	6
Properties or type relations	30	10

IV. THE RECOMMENDER SYSTEM

Accumulating a relevant dataset itself being a research and engineer task (e.g., getting venues basic data from Foursquare), its processing requires thorough development of techniques for all recommender system components available to process a huge amount of information on the fly. Tourist agenda composer meets the requirement of providing real-time services. For routing, it is necessary to find solutions to the problem of the aggregated dataset processing interfaced with map APIs. Aforesaid resulted in several project decisions discussed in this chapter, in particular a recommender function taking submitted preferences into account to provide relevant content.

With the dataset collected by means of reviews analysis, the system should weigh the venue alternatives with user preferences to compose itineraries satisfying the restrictions imposed, and to advise the most optimal according to the recommender function. Recommender strategies could be implemented as follows.

Content-based systems deal with user tastes profiles based on one's ratings. Generally, when creating a profile, a survey is urgent for getting initial information in order to avoid the new-user problem [6].

Case-based systems implement a particular style of content-based ones, undermining the apparent inability of most systems to consider preferences varying over time. New problems are solved by retrieving a case whose specification is similar to the current target problem and then adapting its solution to fit the target [34], [5].

ISSN 2395-8618

Collaborative filtering attributes users to groups with similar preferences within: user-based approach or item-based approach [29].

Hybrid recommender approaches [1].

The rating approach is the one avoided in this research: ratings are widely spread over sites, but apparently not always exact they are, the task of verifying and attributing ratings requires a separate study. Moreover, user preferences differ, which is not reflected by ratings on the whole. Hence, the recommender system is venues-oriented.

A hybrid of the first two strategies is of interest with content-based filtering and implementing some predefined cases (e.g., the must-see sights for first-time visitors to the city). The cold start problem [35] is solved with current preferences indicated for each route query (blocking/non-blocking filtering), with few additional cases possible (extracted from check-ins frequency or manually by experts). With aggregating initial user behavior, detecting and further specifying this very user's modus operandi is subject for the collaborative filter extension of the strategy chosen, subject to design.

The recommender system developed generates a set of routes (itineraries); its inputs for content-based filtering are as follows: (1) an ordered set of venue categories of user's interest; (2) filters for each category (by aspect); (3) an overall time filter.

With venues represented as a graph, a combinatorial optimization problem is solved by means of ant colony optimization technique, which results in suboptimal solution finding (actually, a set of solutions) in a finite time and allows embedding local search. A heuristic is proposed for estimating found routes' costs. As the prototype developed solves a problem of finding the shortest path through categories of objects of interest (e.g., museum + museum + restaurant + museum) with filtering by categories' aspects and considering time restrictions (lower and/or upper bounds), a recommendatory function (RF) is an important part of it. RF penalizes a route for every filter-parameters transgressing by a value between (0, 1]. While some aspects might be absent for a specific object, their penalties are subject to customize in each query. A heuristic of path cost between two nodes is a sum of transfer time and time spent in the node, divided by all the penalties (by filter and by time restriction) multiplication. The overall route cost is a sum of such node-to-node costs, divided by all of the penalties multiplication; then the minimization problem is solved. Dijkstra algorithm is used for routes finding from current node to the next category objects. It is optimized algorithmically to increase performance. Additionally, reducing map nodes, which contain no objects, resulted in x50 acceleration of the well-optimized system.

Beyond the recommender engine, its user interface is subject to implement. The query interface should contain numerical and categorical restrictions for aspects and time, as well as thematic tags (instead of trackbars pattern implemented in [37]) necessary for thematic-focused itineraries forming task. Visual representation of an itinerary might use Gant diagrams in addition to map-based route with links to extracted venues description and/or sites, or otherwise follow the schedule pattern (fully designed in [37]). The engine implemented allows fast creating 10 itineraries per query, and it is easy to provide such feature as recalculating from the current location in case the tourist has changed plans with a time shift. Taking the location factor into consideration is promising for tourist recommender systems, in particular to obtain updated sightseeing information [30] for fast in-place replanning. Location might also be useful for developing an extra widget for creating situational hints (e.g., when a historical building is approached) [8], [19] according to one's tastes/query.

These foresaid project decisions allowed developing a real-time recommender system which forms itineraries with routes satisfying the imposed restrictions, arising from user queries. Optimizations resulted in reduction of this system's recall time, which shifts the system towards production and, in particular, makes it big data-ready, which is actual for big cities and, furthermore, regions (e.g. Provence).

Let the sample input data include 5 categories of venues, the time restriction given not less than 5 hours, starting at 10:00. In Table 5 two routes are provided for this query, each venue accompanied with a timestamp, approximate visit duration and options affected by the query. For this sample, two restaurant positions require simple parameters matching, while the third one and the cultural sites are provided with hashtags marking desired thematic. In case of absence of the 'Lights of Moscow' museum, the recommender system selects a substitute also dedicated to lighting: 'The Ray' cultural center (not a museum, the category is changed). This further implies changes in adjacent positions (the café and the restaurant) to fulfill the route.

TABLE V
SAMPLE QUERY AND ROUTES

Query	Route A	Route B
A café with lunch	10:38 Belucci café (20 min): brunch, dinner, lunch	10:17 Emelya café (15 min): dinner, lunch
A museum (#lighting)	11:01 'Lights of Moscow' museum (56 min)	11:15 'The Ray' cultural center (60 min)
A restaurant with dinner, full bar, wi-fi, outdoor seating, live music	12:38 'The Birch Chalet' restaurant (30 min): full bar, cocktails, live music, outdoor seating, serving lunch, brunch, dinner, with wi-fi	12:37 'Spices and Pleasures' restaurant (35 min): full bar, cocktails, live music, outdoor seating, serving lunch, brunch, dinner, with wi-fi
A cafeteria (#donuts)	13:38 (B: 13:40) The Mega Foods canteen (45 min): breakfasts, desserts, dinner and lunch	
A historic site (#photography)	15:00 (B: 15:02) The Young Photographer Memorial (20 min)	

ISSN 2395-8618

V. Conclusion

The lightweight ontology is described, which covers the domains of restaurants and museums. With this basis the aspects mining method is discussed in detail. Annotated venues are source for automated routes planning and recommending methods, with lightweight ontology refining given heuristics of user's interest in objects. Heuristics for the recommender system and those for estimating objects' relevance to the query consist of rules deriving from lightweight ontology relations over the aspects mentioned. A recommender system for thematic itineraries is designed, which is big data-ready and optimized to allow real-time advising.

Venue aspects aggregated are processed along with user preferences by the flexible route recommending system which generates thematic itineraries throughout the city. A number of such itineraries are generated to user selection, each consisting of particular venues selected by means of recommender techniques, and accompanied with a schedule and a map-based route.

The further development will include constructing different types of relations in order to allow detailed venues analysis. For instance, semantic matching methods should be useful [12], [13] for the "match" relation in case of implementing different overlapping hierarchies, intended for processing detailed information while selecting venues to fulfill the itinerary suiting the request.

References

[1] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. on Knowl. and Data Eng.*, vol. 17, pp. 734–749, June 2005. Washington: IEEE Computer Society.

[2] M. Alexandrov, A. Gelbukh and P. Rosso, "An Approach to Clustering Abstracts", *LNCS* 3513, pp. 275–285, 2005. Berlin: Springer.

[3] M. K. Bergman. (2009, November 23). A reference guide to ontology best practices. In: AI3: Adaptive Information [Online]. Available: http://www.mkbergman.com

[4] M. K. Bergman. (2010, September 13). A new methodology for building lightweight, domain ontologies. In: AI3: Adaptive Information [Online]. Available: http://www.mkbergman.com

[5] D. Bridge, M. Goker, L. McGinty and B. Smyth, "Case–based recommender systems," *Knowledge Engineering Review*, vol. 20 (3), pp. 315–320, 2006. New York: Cambridge University Press.

[6] L. Candillier, K. Jack, F. Fessant and F. Meyer, "State-of-the-art recommender systems," *Collaborative and Social Information Retrieval and Access-Techniques for Improved User Modeling*, pp. 1–22, 2009. Hershey: IGI Global.

[7] P. Cimiano, *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Berlin: Springer (2006)

[8] A. A. Economides, "Requirements of mobile learning applications," *Int. J. of Innovation and Learning*, vol. 5 (5), pp. 457–479, 2009. Geneva: Inderscience Publishers.

[9] Foursquare [Online]. Available: http://www.foursquare.com

[10] A. Gelbukh, G. Sidorov and A. Guzmán-Arenas, "A method of describing document contents through topic selection," in *Proc. of the String Processing and Information Retrieval Symposium and International Workshop on Groupware*, pp. 73–80, 1999. Los Alamitos: IEEE.

[11] F. Giunchiglia, M. Marchese and I. Zaihrayeu, "Encoding classifications into lightweight ontologies", University of Trento, Italy, Technical Report DIT-06-016, March 2006.

[12] F. Giunchiglia and P. Shvaiko, "Semantic Matching," *The Knowledge Engineering Review Journal*, vol. 18 (3), pp. 265–280, 2004. New York: Cambridge University Press.

[13] F. Giunchiglia, P. Shvaiko and M. Yatskevich, "S-match: An algorithm and an implementation of semantic matching," in *Proceedings of ESWS'04, LNCS* 3053, pp. 61–75, 2004. Heidelberg: Springer-Verlag.

[14] M. Gruninger and J. Lee, "Ontology Applications and Design," *Communications of the ACM*, vol. 45 (2), pp. 39–41, 2002. New York: ACM.

[15] N. Guarino, "Formal Ontology and Information Systems," in *Proceedings of Formal Ontologies in Information Systems*, pp. 3–15, 1998. Amsterdam: IOS Press.

[16] J. Hendler, "On beyond ontology", Keynote talk, Second International Semantic Web Conference, Sanibel Island, Florida, USA, 2003, unpublished.

[17] Y.-M. Huang, C.-H. Liu, C.-Y. Lee and Y.-M. Huang, "Designing a Personalized Guide Recommendation System to Mitigate Information Overload in Museum Learning," *Journal of Educational Technology & Society*, vol. 15 (4), pp. 150–166, 2012. International Forum of Educational Technology & Society.

[18] Iknow.travel [Online]. Available: http://www.iknow.travel

[19] I. Keller and E. Viennet, "Recommender Systems for Museums: Evaluation on a Real Dataset," in *IMMM 2015: The Fifth International Conference on Advances in Information Mining and Management*, pp. 65–71, 2015. IARIA.

[20] Y. Kiselev, A. Krizhanovsky, P. Braslavski, I. Menshikov, M. Mukhin and N. Krizhanovskaya, "Russian Lexicographic Landscape: a Tale of 12 Dictionaries," in *Computational Linguistics and Intellectual Technologies: papers from the Annual International Conference "Dialogue" (Moscow, 27-30 May 2015)*. Issue 14, vol. 1, pp. 254–271, 2015. Moscow: RSUH.

[21] T. Kuflik, E. Minkov and K. Kahanov, "Graph-based Recommendation in the Museum," in *CEUR Workshop Proceedings*, vol. 1278. Proceedings of the First International Workshop on Decision Making and Recommender Systems (DMRS2014, Bolzano, Italy, September 18–19, 2014), pp. 46–48, 2014.

[22] D. Lande, A. Snarskii, E. Yagunova, E. Pronoza and S. Volskaya, "Network of Natural Terms Hierarchy as a Lightweight Ontology," in *Thirteenth Mexican International Conference on Artificial Intelligence MICAI 2014*, Tuxtla Gutiérrez, Mexico, 16–22 November 2014. Special session. Revised papers. Gelbukh, A., Espinoza, F. C., Galicia-Haro, S. N. (Eds.), pp. 16–23, 2014. Los Alamitos: IEEE.

[23] N. V. Lukashevich, B. V. Dobrov and D. S. Chuyko, "Selecting word phrases for an automatic text processing system dictionary" (in Russian), in *Computational linguistics and intellectual technologies: Proceedings of Int. Conf. «Dialog–2008»*, pp. 339–344, 2008. Moscow: RSUH.

[24] A. Maslennikova and E. Yagunova, "Information extraction and opinion mining for reviews on the most prominent museums in Russian and English. Methodic and preliminary results," in *New information technologies in automated systems: proceedings of 19th scientific and practical seminar* (in Russian), pp. 68–74, 2016. Moscow: V. M. Keldysh Institute for Applied Mathematics Press.

[25] *OWL Web Ontology Language Guide, W3C Recommendation*, M. K. Smith, C. Welty and D. L. McGuinness (Eds.), 10 February 2004 [Online]. Available: http://www.w3.org/TR/2004/REC-owl-guide-20040210/

[26] E. Pronoza, S. Volskaya and E. Yagunova, "Corpus-based Information Extraction and Opinion Mining for the Restaurant Recommendation System," in *Proceedings of the 2nd Statistical Language and Speech Processing*. L. Besacier et al. (Eds.): SLSP 2014, LNAI, vol. 8791, pp. 272–284, 2014. Berlin: Springer.

ISSN 2395-8618

[27] E. Pronoza, E. Yagunova and A. Lyashin, "Restaurant Information Extraction for the Recommendation System," in *Proceedings of the 6th Language Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, 2nd Workshop on Social and Algorithmic Issues in Business Support: "Knowledge Hidden in Text"*, 2013. Berlin: Springer.

[28] E. Pronoza, E. Yagunova, S. Volskaya and A. Lyashin, "Restaurant Information Extraction (Including Opinion Mining Elements) for the Recommendation System," in *13th Mexican International Conference on Artificial Intelligence, MICAI2014*, Tuxtla Gutiérrez, Mexico, November 16–22, 2014. Gelbukh, A., Espinoza, F. C., Galicia-Haro, S. N. (Eds.). Proc., part I, pp. 201–220, 2014. New York: Springer.

[29] P. Resnick and H. R. Varian, "Recommender systems," *Commun. ACM*, vol. 40 (3), pp. 56–58, March 1997. New York: ACM.

[30] M. K. Sarkaleh, M. Mahdavi and M. Baniardalan, "Designing a tourism recommender system based on location, mobile device and user features in museum," *Int. J. of Managing Information Technology*, vol. 4 (2), pp. 12–21, 2012. Geneva: Inderscience Publishers.

[31] SciKit machine learning library for Python [Online]. Available: http://scikit-learn.org

[32] L. Sebastia, I. Garcia, E. Onaindia and C. Guzman, "E-Tourism: a tourist recommendation and planning application," *International Journal on Artificial Intelligence Tools*, vol. 18 (05), pp. 717–738, 2009. Singapore: World Scientific Publishing Co. Pte. Ltd.

[33] G. Sidorov, A. Gelbukh, H. Gómez-Adorno and D. Pinto, "Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model". *Computación y Sistemas*, vol. 18 (3), pp. 491–504, 2014.

[34] B. Smyth, "Case-based recommender," *The adaptive web*, pp. 342–376, 2007. Berlin, Heidelberg: Springer-Verlag

[35] M. M. Tokareva, L. L. Volkova and A. P. o. Abdullaev, "On a recommender system for itineraries based on user preferences evaluation," in *New information technologies in automated systems: proceedings of 19th scientific and practical seminar* (in Russian), pp. 75–80, 2016. Moscow: V. M. Keldysh Institute for Applied Mathematics Press.

[36] Travel2Moscow [Online]. Available: http://www.travel2moscow.com

[37] Triplantica [Online]. Available: http://www.triplantica.com

[38] Triptomatic [Online]. Available:, http://www.triptomatic.com

[39] M. Uschold and M. Gruninger, "Ontologies and semantics for seamless connectivity," in *SIGMOD Rec.*, 33(4), pp. 58–64, 2004. New York: ACM.

[40] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann and S. Hübner, "Ontology-Based Integration of Information – A Survey of Existing Approaches," in *Gómez Pérez, A., Gruninger, M., Stuckenschmidt, H., Uschold, M. (eds.). Proc. of IJCAI-01 Workshop: Ontologies and Information Sharing*, Seattle, WA, pp. 108–117, 2001.