Revista Romana de Interactiune Om-Calculator 10 (4) 2017, 299-330

Romanian tourism portal analysis through natural language processing, network science and graph database methods

Alex Becheru, Costin Badica

Faculty of Automation, Computers & Electronics, University of Craiova,
Blvd. Decebal, Nr. 107, 200440, Craiova, Romania
E-mail: becheru@gmail.com, cbadica@software.ucv.ro

Abstract. The goal of this study is to lead to a better understanding of the Romanian online tourism phenomenon through the analysis of a representative online portal, i.e. amfostacolo.ro.. The novelty of this study comes from the introduction of complex analysis pipelines based on method gathered from modern computational research fields including natural language processing, network science and graph databases. The study will reveal valuable knowledge on the social behavior and touristic interests of the portal's users as well as future directions of research. Furthermore, we show that the effectiveness of our proposed analysis pipelines is supported by the relevance and nature of the results obtained.

Keywords: online tourism, natural language processing, network science, graph database.

1. Introduction

As during the past decades, the Internet and its functionalities have rapidly expanded and developed to become ubiquitous, so the online global and national (Romanian) tourism has developed and aligned to the current networking advancements. Presently, complex tourism services, like for example: travel, accommodation, and ticketing, can be accessed and combined using Internet. Furthermore, the Internet provides information about tourism related businesses, as well as easier ways to collect feedback on their services and image. This in turn leads to the improvement of their quality. Nevertheless, tourists can take informed decisions with significantly less effort, due to the wide access to information that is available online.

Büyüközkan & Ergün (2011) introduced the term of *smart tourism* to describe "the application of information and communication technologies to the tourism section". A plethora of *smart tourism* portals are available globally providing different functionalities such as: post-visit experiences

sharing, description of tourist attractions, highlights and advices, attraction recommendations, geo-tagged photos, etc. Moreover, such portals address various touristic aspects, like for example: historical places, landscape, sightseeing, food, restaurants, shopping, entertainment, weather, etc. Hence, we can claim that the world of *smart tourism* is worth investigating for gathering its better understanding that might lead to further development of the field.

As object of study, we have chosen the portal *AmFostAcolo*[7]*,* which is one of the most popular *smart tourism* portals in Romania. This portal allows registered users to exchange knowledge by posting impressions and questions. At its core, the portal provides forum like functionalities and targets Romanian language speakers that search touristic information about national and international touristic entities.

Throughout the paper, we shall use the term *touristic entity* to designate a geographical area or geographical attraction of touristic interest. There are no limitations on the area size of an entity, e.g., a touristic entity can refer to *Italy* as a country, *Athens* as a city or to *Manneken Pis as an attraction.* Furthermore, there is no limitation on the touristic entity as it could refer to a city, a cliff, a field of tulips, a hotel, an entertainment park, a statue, etc.

This article is an extension of our work, presented in our previous papers:

- Bădica C. et al. (2014), where we introduced the data extraction mechanism, we made some preliminary analysis and we proposed a sentiment analysis algorithm;
- Becheru A. et al. (2015 July) and Becheru A. et al (2015, September, a), where we presented our results on conducting a Network Science analysis on a touristic content sharing network;
- Becheru A. et al. (2015, September, b), where we proposed a new keyword extraction mechanism from touristic reviews based on network science methods;
- Becheru A. et al. (2016), where we investigated the traits of the communities involved in touristic content sharing.

We structured this paper as follows. Section 2 addresses smart tourism's background based on relevant research papers. The following sections briefly introduce the field of network science and discuss its possible applications in tourism. Section 4 briefly introduces the field of natural language processing and its use in touristic applications. Next, we discuss about graph databases

[7] http://amfostacolo.ro/

and their advantages, also mentioning projects sharing similar goals with our own goals. Section 6 describes the experimental scenarios held, while section 7 discusses the obtained results. Last, we state our conclusions and present possible future work.

2. Smart tourism background

In the early days of the *World Wide Web*, Poon A. (1993) emphasized that information and communication technologies (ICT) will shape the ways of doing tourism. The author claimed that a new type of tourism was emerging driven by "new consumers, new technologies, new production practices, new management techniques and changes in the industry frame condition". According to Carter & Bédard (2001), the World Tourism Organization[8] recognized that ICT was already playing a key role in the entire tourism industry and it published a guideline for the adoption of ICT by destinations and business. By the year 2008, Buhalis & Law (2008) claimed and provided evidence that "the technological revolution experienced through the development of internet has changed dramatically the market conditions for tourism organizations". Moreover, the authors emphasized that ICT technologies have empowered a new type of traveller that is "seeking exceptional value for money and time" and is "keener to pursue their own preferences and schedule". An example of a technological empowered accommodation business is *Airbnb,*[9] it has become the largest hotel chain in the world without actually owning a hotel. This type of business has pushed previous industry leaders, e.g. *Marriott*, to invest heavily in ICT solutions[10]. Furthermore, Airbnb's impact has been noteworthy even for large touristic destinations, such as the city of Barcelona in Spain, see (Gutierrez J. et al., 2016). This work has shown that Airbnb's accommodation created a clear demographical pressure in the city. Furthermore, the results "suggest that Airbnb benefits in greater measure than hotels from proximity of the most visited places in the city".

Later, the researchers have set more focus on developing Smart Tourism. In particular, the development of recommender systems, as well as the use of

[8] http://www2.unwto.org/
[9] https://www.airbnb.com/
[10] http://www.businessinsider.com/marriott-wants-to-be-the-next-airbnb-2017-9

mobile technologies has attracted a lot of scientific work. Borràs J. (2014) has identified the main advances in the use of recommender systems for tourism. Rodriguez-Sanchez, M. C. et al. (2013) and Gavalas D. et al. (2014) have proposed and presented relevant mobile technologies for Smart Tourism. Semantic methods were also used to develop Smart Tourism functionalities, see (Al-Hassan, M. et al., 2015) and (García-Crespo, Á. et al., 2011). Among other research projects, we mention Jiang K.'s (2013) effort to use geo-tagged photos in ranking personalized tourism attractions, as well as investigations in the use of social media applications by Gretzel U. (2015) and No E. & Kim, J. K. (2015). Sigala, M. & Chalkiti K. (2014) investigated the use of Web 2.0. by Greek touristic companies. Although many companies used content sharing networks, few of them effectively exploited this new opportunity for knowledge management activities. Hence, a contribution of our study is to provide an innovative analysis of a touristic content sharing portal, aiming to get a better insight into the potential, opportunities and problems of employing such a business model in the tourism industry.

3. Natural language processing

Texts expressed in natural language are the most common way by which consumers of touristic services present information on Smart Tourism portals. These texts describe users' opinions and/or experiences on various touristic entities. Hence, a need arises to analyze these texts using natural language processing (NLP) methods.

NLP is an applied computing field interested in the automated computational analysis of natural language texts (Manning C. D. & Schütze H., 1999). Jackson, P. & Moulinier I. (2007) present a survey of the use of NLP for online applications. Natural language texts are composed of sentences, hence various linguistic theories were proposed along the ages to understand natural language sentences. They all agree that the foundation of sentence structure relies on identifying and analyzing the various relations among words (subject, complement) (Hristea F. T., 2012). The most common NLP approach of representing these relations is the dependency graph, encapsulating both syntactic and semantic features. Dependency parsers are used to construct dependency graphs, based on the theoretical foundations of dependency grammars, see (Kübler S. et al., 2009). Hence, graphs represent a natural approach to capture and process connections between words in

natural language texts. They were proven to be efficient for lexical acquisition by Widdows D. & Dorow B. (2002), sentence parsing Radev D. R. & Mihalcea R. (2008), word sense disambiguation Navigli R. & Lapata, M. (2010), or text summarization Mihalcea, R. (2004).

A series of research papers supports the utility of NLP within the context of Smart Tourism. Yuan. H et al. (2016) proposed a method to detect travel routes by determining co-occurrences of travel destinations extracted from texts posted on travel blogs, aiming to provide tourists with better travel scheduling. Through aspect-based opinion mining based on *TripAdvisor*[11] Smart Tourism Portal, Marrese-Taylor E. et al. (2013) were effective in determining the sentiment orientation of opinions. García-Pablos A. et al. (2016) introduced a platform (OpeNER) applied to the hospitality domain that processes customer-generated content to obtain valuable information. In another research paper, Filieri R. & McLeay F. (2014) identified strong predictors for the travellers' adoption of information, from online expressed touristic reviews. Hence, travellers could be informed upfront on the quality of the reviews to further simplify their decision making process.

In this study, we rely on NLP methods such as sentence segmentation, tokenization, Part-Of-Speech tagging (POS) and lemmatization. More details of these methods are presented in (Simionescu R., 2011). For benchmarking regarding the accuracy of our proposed approaches, we employ standard measures, like for example *Term Frequency Inverse Document Frequency (tf-idf),* introduced in (Sparck Jones, K., 1972), and initially proposed in the field of information retrieval.

4. Network science

Network Science (NS) or Complex Networks Analysis (CNA) is the research field that studies complex interconnected systems. It is a multidisciplinary research field based on Graph Theory, Computer Science, Physics and Social Sciences. Although there is no clear definition of this field, the majority of researchers agree that its purpose is to investigate non-trivial features of graph problems, not usually addressed by lattice theory or random graphs, i.e. features that are frequently present in natural phenomena. This complexity is

[11] https://www.tripadvisor.com/

caused either by partly regular and random, as well as overlapping and independent phenomena, by the large size of the graph or by the interactions of both. The field has emerged through the study of random graphs by Erdos P. & Rényi A. (1960) and continued with the discovery of the strength of weak ties by Granovetter M. S. (1973). These studies are of most importance as they establish fundamental characteristics of random graphs, respectively the cohesion force that keeps communities together. However, the research field has blossomed after 1990, as computational power become more affordable and data become available thanks to the Internet. Hence, Barabási, A. L. & Albert R. (1999) and Watts D. J. & Strogatz S. H. (1998) were able to explain the occurrence of natural features of the physical world by studying the Internet. The ultimate scope of NS is to understand networks not just as topological objects, but also as frameworks for building complex and distributed dynamical systems, see Newman M. et al. (2011).

Network science was successfully applied in many endeavors, out of which we only mention here some of the most relevant. Barabási A. L. (2011) mentions the use of NS in medicine, especially in neuroscience and in diseases' spread. NS proven of great utility in national security, to discover wanted individuals, see Wilson C. (2010). In the corporate business, information exchange among members of an organization could be improved thanks of applying NS methods, see the book of Cross, R. L. et al. (2010) on organizational network analysis. Costa L. D. F. et al. (2011) described other real-world applications of NS.

NS was also used in the context of Smart Tourism. Some of the works, e.g., Bendle L. J. & Patterson I. (2008), Miguéns J. I. L. & Mendes J. F. F. (2008) and Presenza A. & Cipollina M. (2009), analyzed only the topological aspects of the networks constructed from touristic oriented sources. Results confirm the complex nature of the networks as they have a power-law degree distribution. From a business point of view, some studies were focused on identifying the most relevant actors involved in using and supporting various touristic destinations around the world, see the study on the Elba region in Italy by Baggio R. & Cooper C. (2010) and Waitomo Caves in New Zealand by Pavlovich, K. (2003). Tourists mobility represents another direction of NS application. For example, Taczanowska K. (2014) focused on understanding the structure and the use of hiking trails. More elaborated details on the applications of NS in Smart Tourism are presented in (Baggio, R., 2017) and (Van der Zee E. & Vanneste, D., 2015). As the latter paper emphasizes, there is still a lot of work to do for proving the value of the use of NS in Smart

Tourism, especially from the tourists' point of view. Hence, our study will try to provide knowledge that will ultimately be in the benefit of the tourists.

There are many software tools for supporting NS applications, including toolkits, as well as libraries and APIs. In our NS analysis, we utilized *Gephi*, a tool developed by Bastian M. et al. (2009) with the general purpose of "visualization and exploration for all kinds of graphs and networks". Also, we employed *Python*'s *NetworkX* package for exploration and analysis of networks developed by Hagberg A. et al. (2013). These tools as well as our experiments rely heavily on foundational graph theory / NS metrics such as *Degree*, *Average Degree, Diameter, PageRank, Modularity, Average Clustering Coefficient, Average Path Length, Diameter* and others similar, which will be introduced in the *Discussion and Results* section. Boldi P. & Vigna S. (2014), Page L. et al. (1999), and Blondel V. D. (2008) present in-depth details on these metrics. The presence of social phenomena such as: i) the *small-world* phenomenon described by Travers J. & Milgram S. (1969); ii) *rich-get-richer/power-law degree distribution* phenomenon overviewed by Adamic L. A. & Huberman B. A. (2000); iii) and *preferential attachment* phenomenon discussed by Newman, M. E. (2001), are of great interest for our study. They provide crucial insight about the dynamics of social networks, as well as the behavior of people involved.

5. Graph databases

"Graph databases leverage complex and dynamic relations in highly connected data to generate insight", and "they represent the best way to represent and query connected data" as stated by Robinson I. (2013). Since our work is centered on discovering textual relations from reviews and touristic social relations, the choice of using graph databases comes natural. Moreover, this type of database has the advantage of being readily available for any *join-like* navigation operation, that we will heavily rely on in discovering information, as opposed to SQL and NoSQL database systems. As tool, we used Neo4j[12], as it is an ACID[13] compliant transactional database with native graph storage and processing. Nevertheless, Neo4j has also been

[12] https://neo4j.com/
[13] Atomicity, Consistency, Isolation, Durability

proposed in other Smart Tourism research papers. For example, Karakostas, B. et al. (2017) have employed Neo4j to develop a knowledge-based system for travel mode recommendation, while Jinyan, C. et al. explored tourist's observations for monitoring environmental, with the help of Neo4j.

6. Experimental scenarios

6.1 Data source

As data source we use the information available on Romanian website *AmFostAcolo* (English translation is "I have been there"), which is one of the most popular tourist opinion sharing online portals in Romania. This portal provides semi-structured data oriented on post-visit tourist reviews about a large variety of locations covering specific aspects of accommodation as well as general impressions about touristic sights. From a geographical point of view, tourist attractions depicted on the website mainly cover Romania and Europe with less data for other continents. This is due to the visitation habits and financial restrictions of native Romanian tourists. The data is hierarchically structured according to the country, region, section and geographical place. For example, *Corte Zeuli B&B*[14] is a place under the country *Italy*, region *Puglia* and section *Bari.*

A potential user of the portal can interact with other users in two ways. First, the user can ask questions and provide answers to other questions. This functionality is especially useful for planning purposes, before travelling to a specific location for which enough suitable survey information is not available. Secondly, users can post tourist impressions related to certain places. Hence, for each place, a multitude of impressions posted by various users is available. Moreover, impressions can be commented using echoes issued by other users, with the goal either of presenting further clarifications or of requesting additional information. Although echoes and questions might apparently share similar functionalities, echoes can only exist in the presence of an impression, whereas questions can be posted unrelated to any impression. For a better understanding of the structure of the portal, please consult Figure 1. The portal automatically associates each user with a rank computed based on the user portfolio and feedback provided on the website,

[14] http://amfostacolo.ro/hotel9.php?d=corte-zeuli-bb--bari&id=21509

including facts like: impressions, uploaded photos, posts, and answers.

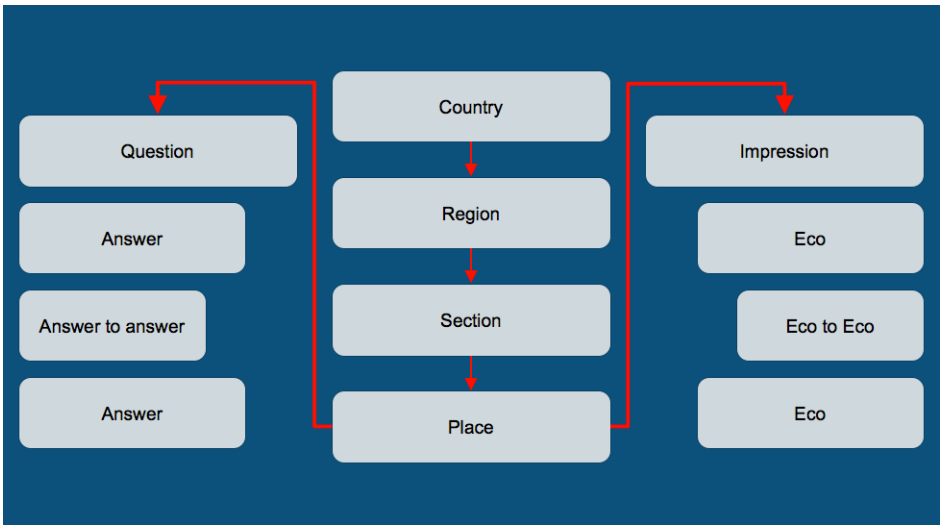

Figure 1. The structure and interaction functionalities of *AmFostAcolo* portal.

6.2 Question-answering scenarios

Our first scenario focuses on the analysis of the *AmFostAcolo* users' social network that emerges from using the *question-answering* functionality. We focus on this functionality for inferring the social network describing users' interaction, as apparently more users have taken advantage of it rather than the *impression-echoing* functionality, i.e. 53172 questions versus 2541 impressions. This difference can be explained as follows. We claim that using the question-answering functionality requires less effort, i.e. reading and understanding of shorter texts, than longer impressions. Also, it is assumed that a user posting an impression on a touristic attraction has already visited it, although the portal does not enforce any kind of verification. Hence, users are limited in posting impressions versus posting questions. Apart from analyzing the resulting social network, we also used Neo4j to link users to questions and answers, with the purpose of developing a method for digesting users' interests.

Our first step was to extract the data from the portal using *web scrapping*[15],

[15] https://en.wikipedia.org/wiki/Web_scraping

more specifically by applying Java's *jsoup*[16] HTML parsing library. We obtained a series of XML files, one for each user, with the XML schema detailed in Figure 2. For experimental purposes and ease of manipulation, we populated a MySQL database with the data present in the XML files (see Figure 3 for the relational schema). Thereafter, we could count 54352 users out of which 3132 have asked at least a question and 7831 that gave at least an answer. In total 8936 (active) users have been using the *question-answering* functionality. Also, we counted 53172 questions and 53150 answers. If we take into consideration only the users that actually posted questions, we get an average of 16.1 questions such user, respectively those that posted answers, we get an average of 6.7 answers per user. Furthermore, we exported the database into *CSV*[17] files, as an intermediary step for importing the data into *Neo4j*. Also, by using MySQL we could easily import the data into *Gephi*[18] network analysis tool. For a better understanding of the entire extraction and pre-processing step, please consult the graphical representation shown in Figure 4.

Graph database experiments

Graph data models consist of nodes and arcs, where nodes represent entities and arcs capture binary relations between entities. Both nodes and arcs can be enriched with attributes, e.g. if a node represents a user we can attach the age attribute. In this scenario, we model three types of entities: *user*, *answer* and *question.* We considered three relations in our model. Relation *r:ASK* connects a *user* (*qUser*) and a *question.* Relation *r:HAS* links a *question* and an *answer.* Relation *r:GIVEN_BY is* defined between an *answer* and a *user* providing that answer. You can inspect the corresponding Neo4j data model in Figure. 5. For querying purposes, we employed *Cypher,* the native querying language for Neo4j. McKnight W. (2014) provides more details on Cypher.

With the help of *AmFostAcolo*'s *question-answering* data available in Neo4j format, we were able to identify the interests of a user, an operation known as *user profiling*. Consider the following scenario: users A, B, and C are a group friends who decided that they should spend their next vacation together. Nevertheless, each user has its particularities and interests and they

[16] https://jsoup.org/
[17] comma separated values
[18] https://gephi.org/

all would like to go to a destination that pleases each of them, a taunting task for the travel agency that they contacted. Having access to their previous *question-answering* data, the travel company can develop a profile for each of them by consistently combining their interests in order to provide them better recommendations.

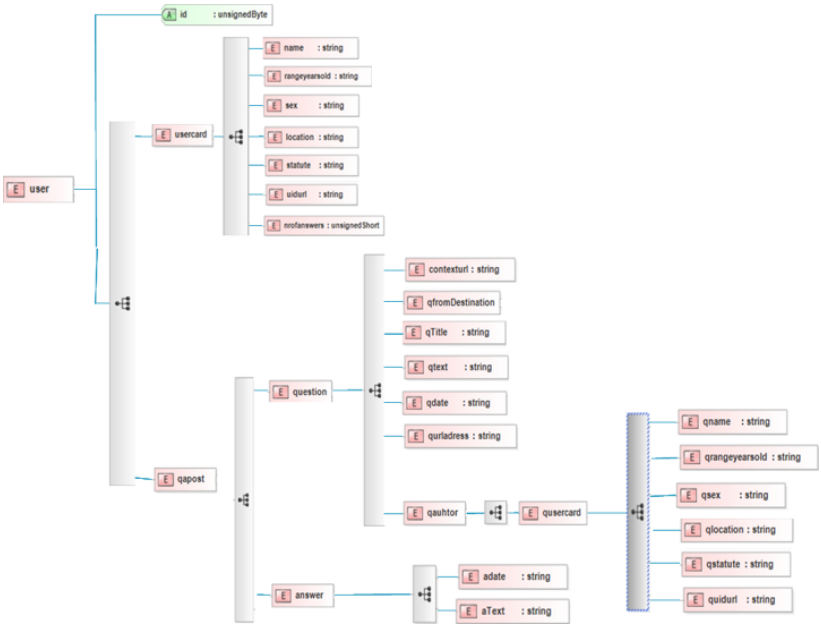

Figure 2. XML schema capturing user interaction data.

To further prove the utility of using Neo4j with the data gathered from the touristic portal consider the following scenario. User A is looking for users that have similar interest, aiming to discover potential new (for user A) touristic locations to visit. Hence, we could compute similarity among users taking in consideration the percentage of questions answered, questions posted by user A, by other users. For example, if user A posted 10 questions out of which 3 were answered by user B and 5 by user C. Then, user A has a similarity of 30% to user B and 50% to user C. In order to answer a user's question another user has to visit the place X on the portal, where the question was addressed. Hence, we can deduce that both users have an interest in place X, from which we deduce the similarity. This scenario can be thought as recommender system for users with similar interests. As we will further

prove, through Neo4j we can easily build this recommender system.

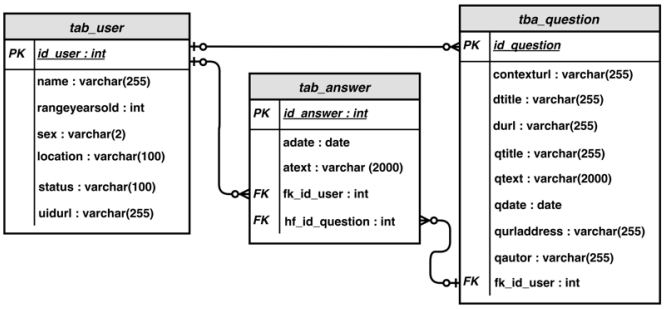

Figure 3. Relational schema describing the extracted data about users, questions, and answers.

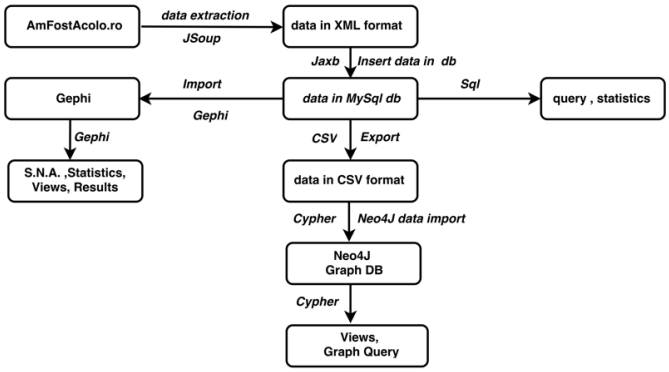

Figure 4. Relational schema of the extracted data.

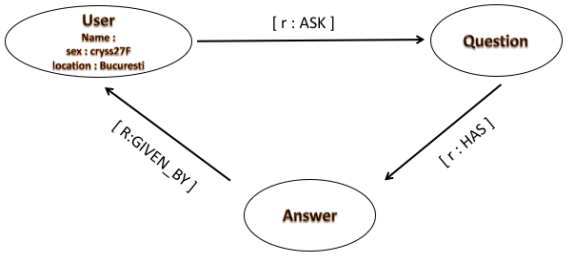

Figure 5. Neo4j data model of users engaged in question-answering interactions.

Network Science experiments

For this part of the study, our entities of interest are the users, represented by nodes of the social graph. We created arcs as follows: an arc connects source

node A to destination node B if node B posted a question upon which node A answered; A is the source node and B the target node. We have considered only the socially engaged users, i.e. users that posted at least one question or one answer. Thus, we created a social network based on the *question-answering* functionality of the website. This network contains 8936 nodes and 29529 arcs. Further, we eliminated self-looping arcs, as they represent answers posted by a user to its own question and those users that answered only to their own questions. Consequently, we reduced our network to 8017 nodes and 25666 arcs. Moreover, we annotated nodes and arcs with their creation date, to enable temporal analysis.

Our current experiments follow three directions:

- Inspect the topological aspects of the network, in order to understand how users socialize.
- Provide a temporal analysis, in order to understand the evolution of the social network.
- Determine inner communities of users and their traits, to identify trends and similitudes among users.

We focused on the following questions that we consider of high interest:

1. Is our network a complex network and if so what type of complex network? What are the traits of the complex network type?
2. Are there present any sociological phenomena and if so what are their traits?
3. Is the network resilient?
4. How did the network evolve over time?
5. Is the network expanding or contracting?
6. Is there any evidence that a specific positive or negative review had or may have significant influence on the community?
7. Can we determine inner-communities and their specific traits?

6.3 Impression-echoing scenarios

This experimental scenario follows the *AmFostAcolo*'s impression-echoing functionality. Using the *web scrapping* technology, we were able to collect impressions posted by 1085 users (423 males and 662 females), totaling a number of 2521 impressions, dispersed over:

- 45 countries

- 161 regions among which 16 country sub-regions and 145 other regions
- 529 sections among which 489 localities (cities, town or villages)
- 1420 tourist locations among which:
    - 534 accommodation units (cottages, pensions, hotels, houses or villas)
    - 886 sections, i.e. representing impressions about a tourist location

In our current scenario, the goal is to define new text analytics methods for analyzing tourist impressions/reviews based on NLP and NS methods, aiming to better capture users' touristic interests. The proposed method is fully unsupervised, see Figure 6 for a graphical representation, and proceeds as follows. For each review, we captured the title, text and metadata. The title and the text were parsed through a chain of NLP methods that include: tokenization, Part-Of-Speech tagging and lemmatization. Hence, the boundaries of each sentence were marked and the part of speech and lemma were attached to each word, in order to unify the representation of all occurrences of the same lexical unit. Next, the review title and content were fully annotated with dependency relations, using the *Romanian Dependency Parser*[19]. A graph/complex network (undirected graph) is resulted, nodes representing lemmas of words and links for dependencies, while keeping the dependency names as link attributes. We used Syntactic filters to remove nodes out of the graph, only keeping nodes with meaningful information. For example, we could consider only nouns and verbs as being meaningful, as carrying relevant semantic information. Hence, we could just keep nodes that represent nouns and verbs and their respective links.

For this experiment, we decided to consider only the words that have stable lexical meanings, i.e. nouns, verbs, adjectives and adverbs. We ignored function words, i.e. prepositions, articles and pronoun, as they only have syntactic functions, i.e. they do not convey valuable information for our analysis. Next, we fusion the different occurrences of the same word in a unique node and then we gather all relations in which each occurrence was involved. The syntactic filtering and unification steps of different occurrences reduced the size of the complex network, making its computational analysis easier, often resulting in a sparse graph/complex network. This method of representing text from reviews/impressions entirely covers all the dependencies, i.e. there was no loss of valuable information. Our analysis

[19] http://nlptools.infoiasi.ro/WebFdgRo/

forked into two directions: *text summarization* using keyword extraction via PageRank algorithm and *words' cohesion checking*.

The first direction assumes the review summarization by identifying relevant words as keywords. We achieved this by computing two different NS metrics for each node: *PageRank coefficient* and *Degree,* and then ranking the nodes according to the metric used. According to *PageRank* metric, a word is a keyword if it is highly linked with words that are also highly linked with other words. According to *Degree* metric*,* a word is considered a keyword if is linked with many other words, with no regards on the others linkage. In order to evaluate the proposed keyword identification methods, we investigate the occurrences of keywords among title words and metadata. The review metadata consists of location information, i.e. country, region, section, and location name. The metadata was also included in this analysis since no words in the title were present in the review's text for 75% of the reviews. This evaluation assumes that discovered keywords and words in the title represent the summary of the review. Moreover, we also used a third method based on the standard *tf-idf* from information retrieval, to compute the keywords capturing the semantics of each review. *We computed* the *term frequency* for each separate review, while the *inverse document frequency* was computed for the entire corpus of review texts. We introduced this third method in order to benchmark our proposed methods versus a standard & widely used keyword extraction method.

The second direction tries to determine if the words defining each review have high cohesion. This would imply that the expressed ideas are tightly connected. This metric could be used to judge if the reviews were not randomly generated, and thus increasing the relevance of our experimental results. We assume that textual cohesion is present in texts if the *small-world* phenomenon is observed in the corresponding graph representations of those texts. This in turn is depicted by evaluating the following NS metrics: *Average Clustering Coefficient, Average Path Length and Diameter.* Moreover, we would expect this to correlate positively with the presence of a *giant component* in the graph that represents the text.

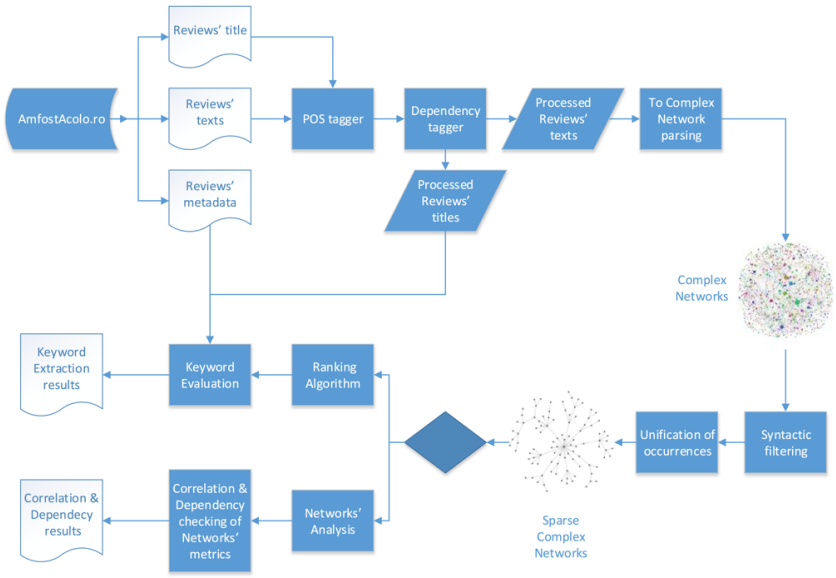

Figure 6. The pipeline of the experiments.

7. Results and discussions

Following the introduction of the experimental details, we are now discussing the results obtained. We have aligned the presentation of the results with the description of each experiment.

7.1 On question-answer scenarios

Graph database experiments

The availability of *AmFostAcolo*'s *question-answering* data in Neo4j format helps in identifying the interests of each user (profiling). As for the first scenario regarding graph databases, a travel agency can use *Cypher* extract the questions and answers of a given user respectively, as depicted below:

```
MATCH (uq:User)-[Ask]->(q:Questions)
   WHERE uq.userId = "15"
   RETURN uq,q
MATCH (uq:User)-[ASK]->(q:Questions) )-[h:HAVE]->
   (a:Answers) - [r:GIVEN_BY] -> [ua:USER]
   WHERE ua.userId = "15"
   RETURN ua,a,q
```

The second graph database scenario, regarding the recommendation of similar users, can be fulfilled by using the *Cypher* query language to determine the sub-graph of users that gave answers to questions asked by a certain user, as follows:

```
MATCH (uq:User)-[Ask]->(q:Questions)-[h:HAVE]->
   (a:Answers) - [r:GIVEN_BY] -> [ua:USER]
   WHERE uq.userId = "15"
   RETURN uq,q,a,ua
```

We acknowledge the fact that many more scenarios can be envisioned and that the data model can be extended. However, we claim that these experiments have proven the usefulness of graph databases in the context of Smart Tourism.

Network science experiments

We start our network science experiments by focusing on the topological aspects supported by various NS metrics. The *Diameter metric* represents the size of the longest path connecting two nodes of the graph, with a computed value of *16* for a graph of 25666 we can argue that observed value is quite low. Furthermore, the *Average Path Length* is *5.042* which we consider to be low for our graph. Both the *Diameter* and the *Average Path Length* are predictors of the *small world* social phenomenon, which we argue to be present in our network due to the relative small values of the metrics in question, see Travers J. & Milgram S. (1969). The implications of the *small world* presence indicate a tighten community of users, where information travels fast. Hence, we can argue that either positive or negative information on a touristic attraction has a significant influence on the community, as it can spread fast and easily from user to user, thus answering question six. Furthermore, additional empirical tests on the diffusion of information strengthen the evidence on the presence of the *small world* phenomenon. In the diffusion experiments, we chose as sources the nodes at the periphery of the network, with a diffusion loss of 70% at each step, assuming that only the neighbors and neighbors of neighbors of the source node could further broadcast. The results have shown that a significant or even large majority of the nodes will receive some information. Figure 7 presents a graphical exemplification of the results of the diffusion experiment.

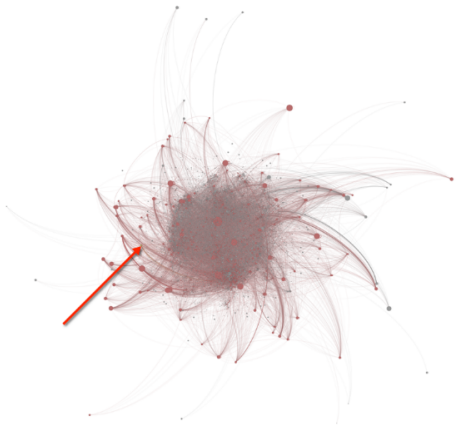

Figure 7. AmFostAcolo's social network with diffusion experiment. The nodes' diameters are proportional to their PageRank coefficient. The color of the nodes depicts if information through the diffusion experiment has reached them (red) or not (grey).

The *Average degree* has a value of *3.201*, while the *Average Weighted Degree* has a value of *5.881*. This means that in average a user responds to 6 questions addressed by 3 other users. Hence, the average social activity is low, which also corroborates with the value of the *Average Clustering Coefficient* of *0.015*. On computing the distribution of the nodes' degree, we observed that the distribution has *scale-free* type. This implies that few users are very active (acting as hubs) while the great majority are not, which is in line with the values obtained for *Degree* and *Average Weighted Degree* metrics. Moreover, this distribution occurs frequently in the Internet, as well as in the natural world, as stated in (Adamic L. A. & Huberman B. A. 2000) and other *Smart Tourism* complex networks as mentioned by Baggio, R. (2017). According to the statistical mechanics of *scale-free* networks studied by Albert R. & Barabási A. L. (2002), this type of networks emerges if both *growth* and *preferential attachment* social phenomena are present. Hence, we found evidence that the social network is growing, answering question 5. Newman M. E. (2001) defined the *preferential attachment* phenomenon as a set of processes in which some quantity, quality, or typically some form of wealth or credit is distributed among a number of individuals or objects according to how much they already have, so those who are already wealthy will receive more than those who are not. This phenomenon is also known as *rich get richer*. In our context this implies that posts from a very active user

(responded to many questions) weighs more (as is more trusted) than an answer from a typical user. The plot presented in Figure 8 shows that the nodes' degree follows a *scale-free* distribution. These results presented on the sociological phenomena provide answers to questions 2 and 4.

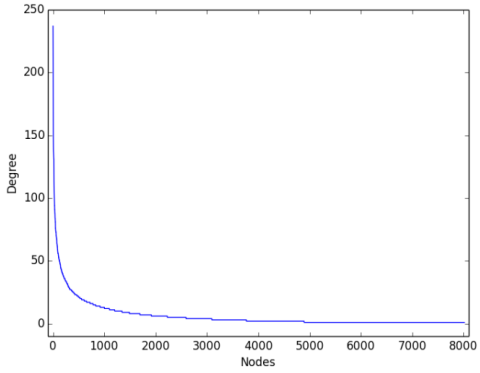

Figure 8. AmFostAcolo's distribution of the nodes' degree for the social network.

According to Figure 7, the visual perception suggests that this network is of the type *core-periphery,* as described in (Hojman D. A. & Szeidl A., 2008). This type of network has a central conglomerate of very well connected (core) nodes, while the outer parts of the network (periphery) are more scattered. To prove the validity of this empirical observation we need to extract the core and analyze its connectivity. Since the core should be the higher connected part of the graph, we could extract it by removing the nodes with a degree less than 30 (10 times the average) and their respective links. Hence, a graph of 300 nodes (3.74% of the total) and 2080 (8.71% of the total links) was obtained (core graph). Then, we computed the following NS metrics and compared the results with those attained for the entire network. Unsurprisingly, the results indicate a higher connectivity of the core graph. The *Average Degree* is *10.267* versus *3.201*, while the *Average Weighted Degree is 28.477* versus *5.881.* These numbers indicate a significantly higher interaction activity with more users, respectively more questions answered in average by the users in the core. The *Diameter* value is *9* versus 16 and the *Average Path Length* value of *3.262* versus *5.042,* implying a tighter connected graph. This also corroborates with the *modularity coefficient* of 0.299 versus *0.48.* Nevertheless, the *Average Clustering Coefficient* depicts

a connectivity that is 4 times larger in the case of the core than on the entire network, *0.068* versus *0.015*. To become part of the core members, one has to be active, hence in a sense we can argue that the social network is meritocratic and can be trusted. The core plays the role of a diffusion enhancer, which is consistent with the rest of our findings. Based on the above findings, we can sustain that the entire network is of *core-periphery* type. A relevant trait for this type of complex network is its resilience to change. This implies that the giant component of the network will be divided only in the case of a catastrophic event, e.g. the loss of a significant part of the most active users. Hence, we consider answered both question 1 and 3.

Next, we focus our topological analysis on the inner communities that rise in the social network (further called communities), thus lowering the granularity level of the network. For the detection of these communities, we used the *modularity* algorithm devised by Blondel, V. D et al. (2008). First we compute the *modularity coefficient* which lies in the range of [-0.5,1]. A positive value of the coefficient suggests that the number of links within groups exceeds the number expected on the basis of chance, i.e. randomly assigned links. For our network, the value of the modularity coefficient is *0.48*. This supports the presence of social communities. The number of detected communities is 146. However, this large number is biased by the presence of small and/or peripheral communities. Hence, we considered only the communities that are part of the graph's giant component. There are 25 such communities. In order to analyse the interactions among these communities, we constructed a directed graph of communities (let's call it *Gco*) as follows. Nodes of the same community fused into a single node, the community node. Links among members of the same community disappeared, while links among members of different communities fused into single links among the representative community nodes' of each user. Based on the initial data from the portal we added directionality to the links, such that the source node is the community node representing the user answering the question and the target is the node representing the community of the user that posted the question. Upon computing the *Average Degree* on the Gco we got a value of 47.2, see Figure 10 for more details. We computed the *Degree* as the sum of the *In-Degree* and the *Out-Degree;* since we have 25 nodes, the maximum value for *In-Degree* and *Out-Degree* is 24. Thus, the maximum value for *Degree* is 48. This implies that users from different communities do interact, which further strengthens the finding that information can travel fast and with ease. Figure 9 presents the graph of communities.

Figure 9. AmFostAcolo's community graph. Each node represents a community; nodes are distinguished by their unique color. The diameter of each node is proportional to its PageRank coefficient.

Since we were able to detect inner-communities in our network, we claim that we have partially answered question 7. To consider this question fully answered, we still have to determine and point out touristic traits that are specific for each community. In this endeavour, we also used the *modularity* algorithm to detect the communities, but we did not restrict its application only to those communities that are part of the network's giant component. In order to capture only the relevant communities, we filtered out those with less than 10 users. We propose two approaches for discovering traits based on the available data, captured from the portal. Concerning users' metadata, we can compute the average scores per community for the following: sex, age, average score given per touristic entity, number of sentences written and rank (computed by the portal). If these averages vary from community to community, then we can assume that they represent specific traits of each community. Since we have obtained the texts for questions and answers for each user, we propose to aggregate these texts per each community and then determine textual traits for each resulting text. First we parsed each users' questions and answers using the *Part-of-Speech* tagger for Romanian language described in (Hristea, F. & Popescu M., 2003), thus obtaining a set of tokens together with their occurrence frequency. Next, we aggregated the tokens for each community of users and computed their community frequency. We filtered tokens to keep only those tokens representing nouns, verbs and adjectives, as the rest of the tokens possess only syntactic information that is not relevant for our purpose. Moreover, we ranked decreasingly by their occurrence frequency the nouns, verbs and adjectives

and considered the top 10 most frequent as candidates for capturing touristic traits. Hence, for each community we obtained 3 lists of the 10 most frequent tokens for the parts of speech considered.

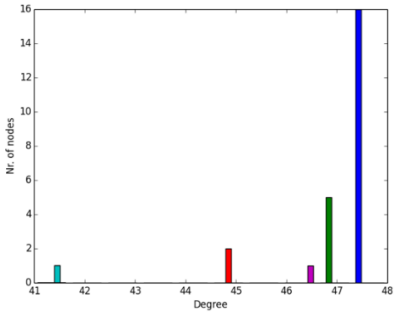

Figure 10. AmFostAcolo's *Degree* distribution for nodes in the graph of communities (*Gco*)

Regarding the metadata traits, we were able to reach some results. The users' age varies between 30 and 36 years old and the users' rank varies between 3.8 to 5.4 (on a scale from 0 to 12), for each community. Due to their low variance, these parameters cannot be considered as representing specific traits, while the users are apparently uniformly distributed. Unfortunately, the results obtained for the other metadata (sex, average score given per touristic entity and number of sentences written) were not significant, so this aspect needs to be further examined in future works. However, the communities have a larger variance for noun and verb textual tokens. Hence, we argue that the top most frequent nouns and verbs can be seen as specific touristic traits for each community. As a fact, we observed that the verb *to eat* is the only frequently used verb for the vast majority of communities. Hence, we can argue that *eating* is considered a key activity regardless of tourists various interests. Despite having also a large variance, the large majority of adjectives refer to monetary values, which do not provide useful information by themselves. Here the context plays a key role that we did not properly consider in this experiment. Hence, we claim that further studies should consider nouns and verbs as specific traits for each community, while adjectives should be captured together with some context information.

Our topological study could continue by further lowering the granularity of the analysis towards nodes. However, we left this kind of deeper investigation as future work. Nevertheless, we identified two directions of interest. The first one should focus on the nodes importance in the social

network. This can be determined using the *PageRank* algorithm or using the *betwenness* metric to highlight the users that play a major role in the spread of information. Rewarding (given by the portal) the users of great social importance may represent the necessary incentive for those users to further continue and/or improve their presence on the portal. Thus, the portal, the users in question and the other users will see the benefits of a more "social" network. Another direction involves using the geographical metadata for the users and questions, to determine touristic trends of visitation. It is possible to create a complex network where nodes represent locations and links represent interest in locations that can be further analysed as described in (Miguéns J. I. L. & Mendes J. F. F., 2008).

As initially planned, we also conducted a temporal analysis of the social network. In order to fulfill the analysis, we retrieved from *AmFostAcolo* portal the date of each question and answer posting. As methodology, we considered capturing a yearly snapshot, from 2010 to 2015 on the 1st of January of each year. Previous data were also available, but the network was of negligible size. As Figure 11 shows, the network continuously grew in size and developed its shape, from the core to the periphery (like a tree). The growth process appears to be cyclical as follows:

1. *Expansion step*. It consists in attracting new users in the network. Observe the years 2010 and 2014.
2. *Development step*. It consists in new links creation among the current users with less attraction of new users. Observe the years 2011, 2012, 2013 and 2015.
3. Return to step 1.

Hence, the cycle appears to span a number of years, with the first year being the expansion step, while the following 3 years are devoted to the development step. However, we believe that this temporal analysis must be repeated for the following years (2016 and afterwards) for strengthening this analysis. Nevertheless, we have yet again brought evidence that the social network is expanding, thus providing an answer to question 5.

7.2 Impression-echoing scenarios

Table 1 shows a summary of the results for the keyword extraction process. Before commenting on the results, we should mention that just for 723 reviews (out of 2521) we could find a common word in the title and in the

text of the review, while considering nouns, adjectives, adverbs, and verbs.

Hence the inclusion of the metadata is fully justified as it leads to the increase of the number of analysed impressions/review to more than 2000 texts in which at least one keyword can be found in the reviews' title or content, as it is shown by experiments 3 and 4 from Table 1. Analyzing these results, we observe with high certainty that each of the *PageRank* or *Degree* methods for computing keywords is better than *tf-idf*. However, we note that the cross-checking of the list of obtained keywords with the reviews' title is not effective, as in most cases titles contain words that more or less summarize the reviews, while often the words occurring in the title are not present in the review content. Moreover, we recommend to use only *nouns* and *adjectives* as possible keywords, as in our experiments they gave the best results with less computational effort.

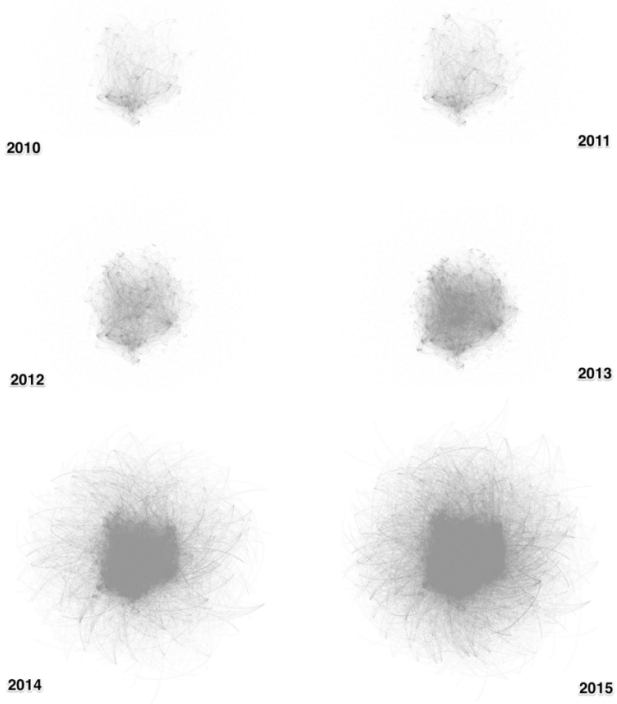

Figure 11. AmFostAcolo's time lapse of the social network as captured of the 1st of January of each year.

Regarding cohesion and word distribution experiments, we were able to observe the *small-world* phenomenon in the majority of the analyzed networks. We support this assumption by observing that the values of the *average clustering* coefficient were higher than of randomly generated networks with the same number vertices. The values of the *Average Path Length* and *Diameter* metrics were relatively low, compared to the number of vertices, as can be noticed in Figure 12 that displays the histogram of the obtained values for the *Diameter* metric. We also observed that approximately 64% of the networks had only one connected component, while the rest had 2 or 3 connected components. Moreover, every network contained a *Giant Component,* containing more than 95% of the nodes. Hence, even if 2 or 3 components were found then one of them was the dominant, while the others were relatively small (counting the number of nodes). For example, Figure 13 and 14 show the graphical representation of a network with 2 respectively 3 connected components. The presence of the *small-world* phenomenon and dominance of *one giant component* confirms that the reviews were not randomly, but rather naturally generated, as they were composed and written by human reviewers. Moreover, we noticed the *scale-free* distribution of the nodes' degree, (nodes representing tokenized words), an observation falling within the general trend of networks acquired from lexical corpora, as mentioned in (Radev D. R. & Mihalcea R., 2008).

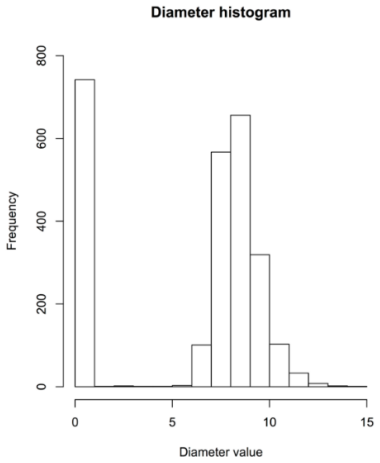

Figure 12. Histogram of the *Diameter* values obtained for the networks defined by reviews.

Table 1. AmFostAcolo's keyword extraction results. The columns *Title* and *Text* indicate the Part-of-Speech of the remaining words, after the semantic filtering step: N for noun, A for Adjective, Ad for adverb and V for verb. The M stands for the inclusion of location metadata. The following columns indicate the percentage of keywords extracted, versus the number of words in the reviews. Hence, the column 20% depicts that after the ranking of the words, keywords are the first 20%. For example, the result on line 5 and column 20% suggests that by using the PageRank algorithm to determine the keywords, considering the top 20% words as keywords, we could find that 2% of keywords were present in the title. The '# reviews' column indicates the number of reviews considered in each experiment.

#	Method	Title	Content	100 %	50%	33 %	20 %	10 %	5%	# reviews
1	Degree	NA	NA	11%	50%	33%	2%	1%	0.6%	2521
2	Degree	NA	NA	43%	18%	11%	8%	4%	2%	703
3	Degree	NAAdV	NAAdV	36%	16%	6%	9%	3%	1%	723
4	Degree	NAAdV M	NAAdV	27%	13%	8%	5%	2%	1%	2184
5	PageRank	NAM	NA	28%	14%	8%	6%	2%	1%	2171
6	PageRank	NAAdV M	NAAdV	27%	13%	8%	5%	2%	1%	2184
7	PageRank	M	NA	36%	17%	11%	7%	3%	1%	2081
8	Tf-idf	NAM	NA	17%	8%	5%	3%	2%	1%	2107

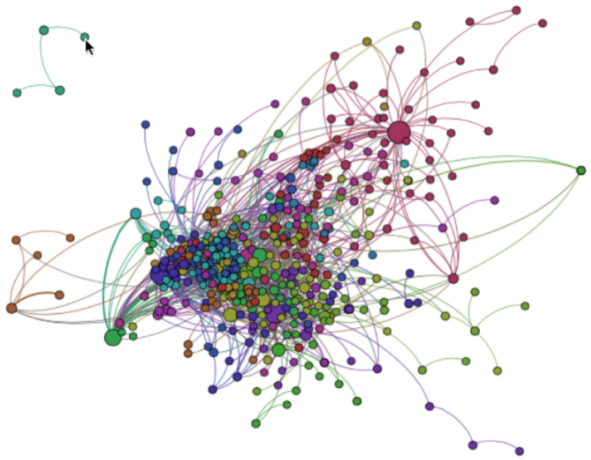

Figure 13. Example of a network with 2 connected components. The radius of each node is proportional to its PageRank coefficient. The color of each node denotes membership to a community, as computed with the *Modularity algorithm.*

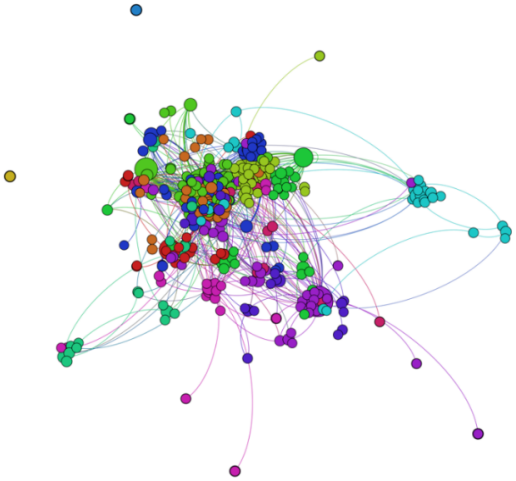

Figure 14. Example of a network with 3 connected components. The radius of each node is proportional to its PageRank coefficient. The color of each node denotes membership to a community, as computed with the *Modularity algorithm.*

8. Conclusions and future work

In this work, we have used several computational methods, techniques and tools stemming from various fields: Network Science, Natural Language Processing, Graph Theory, and Graph Databases, to analyze a portal for sharing touristic opinions (*AmFostAcolo*). The analysis was inspired by the portal functionalities, including question asking and answering, impression and review posting, as well as commenting and echoing. The main goal of the study was to achieve a better understanding of the touristic collaboration phenomenon.

Regarding the data extracted and analyzed from the *question-answering* functionality, we draw a number of conclusions. We have proven the usefulness of using the Neo4j graph database model for retrieval of information that is relevant for two real-world touristic scenarios. The network science experiments allowed us to state that the network is stable and resilient, while continuing a cyclical trend of expansion that attracts new users and then matures their connection. The presence of various social phenomena has shown that the social environment has meritocratic traits, i.e. users with

more intense activity are better trusted while acting as enhancers of information diffusion. Moreover, we have proven that information travels fast and easy even if the source user is peripheral. We were able to identify inner-communities of users and give evidence that the collaboration among the communities is intense. We have also experimented and proposed a NLP-based method to identify touristic traits of each community.

Using the texts extracted from the *impression-echoing* functionality, we proposed a new method, based on NS and NLP that is able to summarize the texts via keyword extraction. The results have shown that our two proposed methods (*Degree* and *PageRank*) perform better than the widely used *tf-idf* information retrieval method. Also, we have brought evidence that nouns and adjectives are the parts of speech carrying out the most relevant touristic information. Nevertheless, the use of location metadata has proven useful for our experiments and we claim that it carries out relevant information. Furthermore, we have applied NS methods to determine the cohesion of the posted reviews and their similarity to real-world texts. This is useful to check that the reviews were not automatically generated by softbots.

As a general conclusion, we have proven that our proposed approaches can bring new valuable insights on how tourists collaborate via touristic portals. However, we acknowledge that this study can be expanded by adding more touristic scenarios and eventually also expanding the data model. A further analysis is required towards inspecting visitation trends, based on previous studies of complex networks. Regarding our proposed keyword extraction methods, we acknowledge the fact that validating the keywords based on words in the title has its caveats, thus other validation methods need to be developed. However, as current results have shown our proposed method gives better results than *tf-idf*.

References

Adamic, L. A., & Huberman, B. A. (2000). Power-law distribution of the world wide web. *Science*, *287*(5461), 2115-2115.

Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, *74*(1), 47.

Al-Hassan, M., Lu, H., & Lu, J. (2015). A semantic enhanced hybrid recommendation approach: A case study of e-Government tourism service recommendation system. *Decision Support Systems*, *72*, 97-109.

Bădică, C., Colhon, M., Şendre, (2014, September). Sentiment analysis of tourist reviews: Data preparation and preliminary results. In *Proceedings of the 10th International*

Conference Linguistic Resources And Tools For Processing The Romanian Language, ConsILR (pp. 135-142).

Baggio, R. (2017). Network science and tourism–the state of the art. *Tourism Review*, *72*(1), 120-131.

Baggio, R., & Cooper, C. (2010). Knowledge transfer in a tourism destination: the effects of a network structure. *The Service Industries Journal*, *30*(10), 1757-1771.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509-512.

Barabási, A. L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, *12*(1), 56-68.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *Icwsm*, *8*, 361-362.

Becheru, A., Bădică, C., & Antonie, M. (2015, July). Towards Social Data Analytics for Smart Tourism: A Network Science Perspective. In *Workshop on Social Media and the Web of Linked Data* (pp. 35-48). Springer International Publishing

Becheru, A., Buşe, F., Colhon, M., & Bădică, C. (2015, September, a). Tourist review analytics using complex networks. In *Proceedings of the 7th Balkan Conference on Informatics Conference* (p. 25). ACM.

Becheru, A., Bădică, C., & Antonie, M. (2015, September, b). Complex Network Analysis of a Tourism Content Sharing Network. In *Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2015 17th International Symposium on*(pp. 407-414). IEEE.

Becheru, A., & Bădică, C. (2016). A deeper perspective of online tourism reviews analysis using natural language processing and complex networks techniques. In *Proceedings of the 11th International Conference Linguistic Resources And Tools For Processing The Romanian Language, ConsILR Editors: Maria Mitrofan Daniela Gîfu Dan Tufiș Dan Cristea*, pp. 189.

Bendle, L. J., & Patterson, I. (2008). Network density, centrality, and communication in a serious leisure social world. *Annals of Leisure Research*, *11*(1-2), 1-19.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Boldi, P., & Vigna, S. (2014). Axioms for centrality. *Internet Mathematics*, *10*(3-4), 222-262.

Borràs, J., Moreno, A., & Valls, A. (2014). Intelligent tourism recommender systems: A survey. *Expert Systems with Applications*, *41*(16), 7370-7389.

Buhalis, D., & Law, R. (2008). Progress in information technology and tourism management: 20 years on and 10 years after the Internet—The state of eTourism research. *Tourism management*, *29*(4), 609-623.

Büyüközkan, G., & Ergün, B. (2011). Intelligent system applications in electronic

tourism. *Expert systems with applications*, *38*(6), 6586-6598.

Carter, R., & Bédard, F. (2001). E-Business for Tourism-Practical Guidelines for Destination and Businesses. *Madrid: World Tourism Organisation*.

Colhon, M., Cerban, M., Becheru, A., & Teodorescu, M. (2016, August). Polarity shifting for Romanian sentiment classification. In *INnovations in Intelligent SysTems and Applications (INISTA), 2016 International Symposium on* (pp. 1-6). IEEE.

Costa, L. D. F., Oliveira Jr, O. N., Travieso, G., Rodrigues, F. A., Villas Boas, P. R., Antiqueira, L., ... & Correa Rocha, L. E. (2011). Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, *60*(3), 329-412.

Cross, R. L., Singer, J., Colella, S., Thomas, R. J., & Silverstone, Y. (2010). The organizational network fieldbook: Best practices, techniques and exercises to drive organizational innovation and performance. John Wiley & Sons.

Erdos, P., & Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, *5*(1), 17-60.

Filieri, R., & McLeay, F. (2014). E-WOM and accommodation: An analysis of the factors that influence travelers' adoption of information from online reviews. *Journal of Travel Research*, *53*(1), 44-57.

García-Crespo, Á., López-Cuadrado, J. L., Colomo-Palacios, R., González-Carrasco, I., & Ruiz-Mezcua, B. (2011). Sem-Fit: A semantic based expert system to provide recommendations in the tourism domain. *Expert systems with applications*, *38*(10), 13310-13319.

García-Pablos, A., Cuadros, M., & Linaza, M. T. (2016). Automatic analysis of textual hotel reviews. *Information Technology & Tourism*, *16*(1), 45-69.

Gavalas, D., Konstantopoulos, C., Mastakas, K., & Pantziou, G. (2014). Mobile recommender systems in tourism. *Journal of network and computer applications*, *39*, 319-333.

Gutierrez, J., Garcia-Palomares, J. C., Romanillos, G., & Salas-Olmedo, M. H. (2016). Airbnb in tourist cities: comparing spatial patterns of hotels and peer-to-peer accommodation. *arXiv preprint arXiv:1606.07138*.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, *78*(6), 1360-1380.

Gretzel, U., Werthner, H., Koo, C., & Lamsfus, C. (2015). Conceptual foundations for understanding smart tourism ecosystems. *Computers in Human Behavior*, *50*, 558-563.

Hagberg, A., Schult, D., Swart, P., Conway, D., Séguin-Charbonneau, L., Ellison, C., ... & Torrents, J. (2013). Networkx. High productivity software for complex networks. *Webová strá nka https://networkx. lanl. gov/wiki*.

Hristea, F. T. (2012). The Naïve Bayes Model for Unsupervised Word Sense Disambiguation: Aspects Concerning Feature Selection. Springer Science & Business Media.

Hristea, F., & Popescu, M. (2003). A dependency grammar approach to syntactic analysis

with special reference to Romanian. *Building Awareness in Language Technology. University of Bucharest Publishing House*.

Hojman, D. A., & Szeidl, A. (2008). Core and periphery in networks. *Journal of Economic Theory*, *139*(1), 295-309.

Jackson, P., & Moulinier, I. (2007). Natural language processing for online applications: Text retrieval, extraction and categorization (Vol. 5). John Benjamins Publishing.

Jiang, K., Yin, H., Wang, P., & Yu, N. (2013). Learning from contextual information of geo-tagged web photos to rank personalized tourism attractions. *Neurocomputing*, *119*, 17-25.

Jinyan, C., Susanne, B., & Bela, S. Citizen science–exploring tourists' observations on micro blogs as a tool to monitor environmental change.

Karakostas, B., & Kardaras, D. K. A (2017) Knowledge Graph for Travel Mode Recommendation and Critiquing.

Kübler, S., McDonald, R., & Nivre, J. (2009). Dependency parsing. *Synthesis Lectures on Human Language Technologies*, *1*(1), 1-127.

Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.

Marrese-Taylor, E., Velásquez, J. D., Bravo-Marquez, F., & Matsuo, Y. (2013). Identifying customer preferences about tourism products using an aspect-based opinion mining approach. *Procedia Computer Science*, *22*, 182-191.

McKnight, W. (2014). Chapter Twelve–Graph databases: when relationships are the data. *Information Management*, 120-131.

Miguéns, J. I. L., & Mendes, J. F. F. (2008). Travel and tourism: Into a complex network. *Physica A: Statistical Mechanics and its Applications*, *387*(12), 2963-2971.

Mihalcea, R. (2004, July). Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*(p. 20). Association for Computational Linguistics.

Navigli, R., & Lapata, M. (2010). An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE transactions on pattern analysis and machine intelligence*, *32*(4), 678-692.

Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical review E*, *64*(2), 025102.

Newman, M., Barabasi, A. L., & Watts, D. J. (2011). *The structure and dynamics of networks*. Princeton University Press.

No, E., & Kim, J. K. (2015). Comparing the attributes of online tourism information sources. *Computers in Human Behavior*, *50*, 564-575.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab.

Pavlovich, K. (2003). The evolution and transformation of a tourism destination network: the Waitomo Caves, New Zealand. *Tourism Management*, *24*(2), 203-216.

Presenza, A., & Cipollina, M. (2009). Analysis of links and features of tourism destination's stakeholders. An empirical investigation of a South Italian Region.

Poon, A. (1993). Tourism, technology and competitive strategies. CAB international.

Radev, D. R., & Mihalcea, R. (2008). Networks and natural language processing. *AI magazine*, *29*(3), 16.

Robinson, I., Webber, J., & Eifrem, E. (2013). *Graph databases*. " O'Reilly Media, Inc.".

Rodriguez-Sanchez, M. C., Martinez-Romo, J., Borromeo, S., & Hernandez-Tamames, J. A. (2013). GAT: Platform for automatic context-aware mobile services for m-tourism. *Expert Systems with applications*, *40*(10), 4154-4163.

Sigala, M., & Chalkiti, K. (2014). Investigating the exploitation of web 2.0 for knowledge management in the Greek tourism industry: An utilisation–importance analysis. *Computers in Human Behavior*, *30*, 800-812.

Simionescu, R. (2011). Hybrid pos tagger. In Proceedings of Language Resources and Tools with Industrial Applications Workshop (Eurolan 2011 Summer School), Cluj-Napoca, Romania (pp. 21-28).

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *28*(1), 11-21.

Taczanowska, K., González, L. M., Garcia-Massó, X., Muhar, A., Brandenburg, C., & Toca-Herrera, J. L. (2014). Evaluating the structure and use of hiking trails in recreational areas using a mixed GPS tracking and graph theory approach. *Applied Geography*, *55*, 184-192.

Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, 425-443.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, *393*(6684), 440-442.

Widdows, D., & Dorow, B. (2002, August). A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1* (pp. 1-7). Association for Computational Linguistics.

Wilson, C. (2010). Searching for saddam: Why social network analysis hasn't led us to osama bin laden. *Slate (February 26, 2010)*.

Yuan, H., Xu, H., Qian, Y., & Li, Y. (2016). Make your travel smarter: Summarizing urban tourism information from massive blog data. *International Journal of Information Management*, *36*(6), 1306-1319.

Van der Zee, E., & Vanneste, D. (2015). Tourism networks unravelled; a review of the literature on networks in tourism management studies. *Tourism Management Perspectives*, *15*, 46-56.