# Tourism Websites Network: Crawling the Italian Webspace

Alessandro Longheu, Giuseppe Mangioni, Marialaura Previti
Dipartimento di Ingegneria Elettrica, Elettronica ed Informatica (DIEEI)
University of Catania
Catania, Italy
e-mail: alessandro.longheu@dieei.unict.it, giuseppe.mangioni@dieei.unict.it, ml.previti@gmail.com

***Abstract*— The relevance of tourism is increasing more and more in the globalized economy. To improve management of tourism activities, the analysis of tourism related data deserves a major attention. In this work, we investigate on the Italian tourism website network, exploiting crawling and classifying techniques in order to extract and analyze such network, with the final goal of providing useful information for tourism management stakeholders.**

***Keywords*— *web pages categorization; text classification; Support Vector Machine (SVM); complex networks; tourism management.***

## I. INTRODUCTION

Tourism is one of the main entries in globalized economy, and is becoming more and more a challenging issue in many research areas even apparently unrelated, as sustainable (tourism) development [1], health-care and well-being [2], as well as many others [3].

Even if the pervasive adoption of mobile devices and related apps for tourism resource discovery and management represent a cutting-edge topic [4], the huge amount of web sites currently available still represents the main data source for tourism information [5].

In this paper, we aim at focusing on the Italian website tourism network in order to analyze such network and its properties. Our final goal is to exploit this information to improve the comprehension about tourism phenomenon, for which a number of models have been proposed in the past with limited success in providing satisfactory insights [6].

Actually, each term of "Italian website tourism network" should be further specified according to topics we want to highlight:

- by "Italian" we mean any website containing information written in Italian language, since we believe that information in other languages may either represent a simple duplicate of the former, or they are not intended for the Italian webspace;
- by "website" we mean the site seen as the atomic unit of information, i.e., we gather (text) information from all pages belonging to the same site, of course, we limit the data amount stopping the of collecting web pages from the same site according a given depth;
- by "tourism" we mean that each site contains tourism related information, we perform the classification using a SVM based text classifier that has been develop using the KNIME software.
- finally, by "network" we mean that we aim at gathering hyperlinked Italian tourism websites. In particular, the crawling process used to build such a network starts from an initial seed made by official, institutional regional tourism websites, further considering all hyperlink contained within each visited page. The process is properly managed to avoid downloading huge amount of data, as explained below.

After the (large) crawling operation cited so far, we started to analyze such data. In this work, we consider in particular one of the first crawled Italian region ("Abruzzo") and we illustrate the related network and its analysis. These preliminary results are promising in terms of their relevance in providing significant information about how many Italian tourism websites are present, how they are connected and finally how such information can be exploited to improve tourism management.

The rest of paper is organized as follows: in Section II, an overview of related work about website classification is introduced, then in Section III we describe our work in more detail. In Section IV first results are presented, finally providing in Section V our conclusions and further works.

## II. WEB CONTENT CLASSIFICATION

Web page classification is the process of assigning a web page to one or more categories. In recent years, several methodologies for the automatic classification of web pages have been introduced. The most common approach is based on web pages information content categorization: images analysis, text classification, metadata extraction, document structure analysis, etc.

To achieve this goal, F. Sebastiani [7] mainly focused on traditional textual classification based on supervised machine learning techniques, C. Lindermann [8] identified and analyzed structural properties, which reflect the functionality of a Web site to divide them into five most relevant functional classes, Z. Xu [9] proposed a statistical web page classification approach, which incorporates heterogeneous data sources extracted from web pages like title, metadata, anchor texts, URLs, links and formulates them into a common format of kernel matrix, P. Calado [10] used link information in combination with content information to improve classification results for web collections, A. Sun [11] proposed the use of Support Vector

 ISBN: 978-1-61208-510-4 

Machine classifiers to classify web pages using both their text and context feature sets.

Although no new techniques are presented in this paper, we used a combination of the aforementioned to analyze the Italian tourist webspace.

Although there are various classifiers that allow to review the English-language texts, until today, no one has modified a classifier for the purpose of classifying large-scale texts in Italian language, so we have had to face this challenge to perform this work.

## III. CRAFTING ITALIAN TOURISM WEBSITE NETWORK

### A. Overview

As introduced in Section I, the work here proposed begins from considering a set of Italian official tourism websites (actually not shown here) that were manually classified.

We used these sites as a seed to start crawling, in particular we considered for each site its home page together with related first level pages (i.e., those directly linked by the home page), extracting all hyperlinks to proceed with crawling process, and retaining the whole text of such pages to further classify the site.

Each website will be a node in the network, and the crawling continues until a specified number of hops is reached, hence the overall network is built. This network may include site with non-italian language text, and/or unrelated to the tourism topic, therefore this is simply the network of (hyper)linked sites reachable from the seed.

In order to extract the subnetwork of Italian tourism websites, we first discard all sites whose language is not Italian, then we perform a classification to establish whether a site is about tourism or not, using the seed as a training set. All nodes that “survive” these filtering phases will compose the Italian tourism (sub)network.

Finally, to cope with crawling time and with the huge amount of data to analyze, we split the whole process over the 20 italian geographical regions. In the following, each step of our proposal is described in more detail.

### B. Filtering Crawled data

After having crawled data, the next step was the filtering of web pages to get only the Italian ones, in fact an initial data analysis showed that most of tourist web pages are part of multilingual sites and are linked to foreign countries pages.

Foreign words can adversely affect the text classification because they can't be prefiltered with preprocessing instruments that use the Italian vocabulary (e.g., stemmers, word filter stop), therefore they are placed in the bag-of-word as they are.

Therefore we developed a whitelist filter used to discard web pages with a top-level domain different from .it, .com, .org, .net, .tv, .info, .eu, .ch, .at and .fr, then we filtered the selected pages using the java library language-detection [12] to further discard the pages with Italian domain, but written in a foreign language. Note that in the whitelist we included top level domain of countries bordering Italy since this geographic proximity sometimes implies that sites are bi-lingual, e.g. a .fr website is likely to include Italian text.

Today many websites are created using scripts, therefore extracting text useful for the classification is difficult and, in some cases, is also insufficient, then we need to preliminarily extract additional textual useful information to be added to the text from the body of the page. In order to do this, our filter not only performs the removal of tags and scripts, but preliminarily examines tags containing the metadata (HTML and OG [13]) of title, description and keyword and, after text cleaning, adds this information on top of the extracted text.

### C. Classification tool

After a practical comparison of six most used free software tools for general data mining today available and thanks the Jovic [14] description of algorithms and procedures supported by these tools, we decide to use KNIME [15], an open-source general-purpose data mining tool based on dataflow architecture. This tool offers several configurable building blocks in the core installations and various extensions, including "text processing" and "mining" plugins that we have used to achieve our purpose.

### D. Training set

Among the classification techniques analyzed [16][17], we decided to use Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO) [18], a simple algorithm that quickly solves the SVM quadratic programming problem without any extra matrix storage and without invoking an iterative numerical routine for each sub-problem, significantly reducing training time.

The first step to use SVM is the creation of a data set, so, starting from a list of tourist web page URLs manually created searching on internet for hotels, travel agencies, tourist destinations, etc., with our filter we automatically extract text and save it in textual.

After this we have assessed the results and have discarded texts with insufficient size or content. The remaining texts became part of the tourist texts folder.

Similar work has been done to create non-tourist texts folder.

To evaluate the quality of the work performed, we created a KNIME program to do cross-validation (Fig. 1), i.e., to partition data into two segments: one used to train the model and the other used to validate it. In typical cross-validation, the training and validation sets must cross-over in successive rounds such that each data point has a chance of being validated against. We therefore repeated ten times the classification and produced an error rate table shown in Fig. 2, where during each cycle 9 folds are used as training set and the remaining one is the test set.

The results show that the average of classification accuracy is 89,75%.

 ISBN: 978-1-61208-510-4 

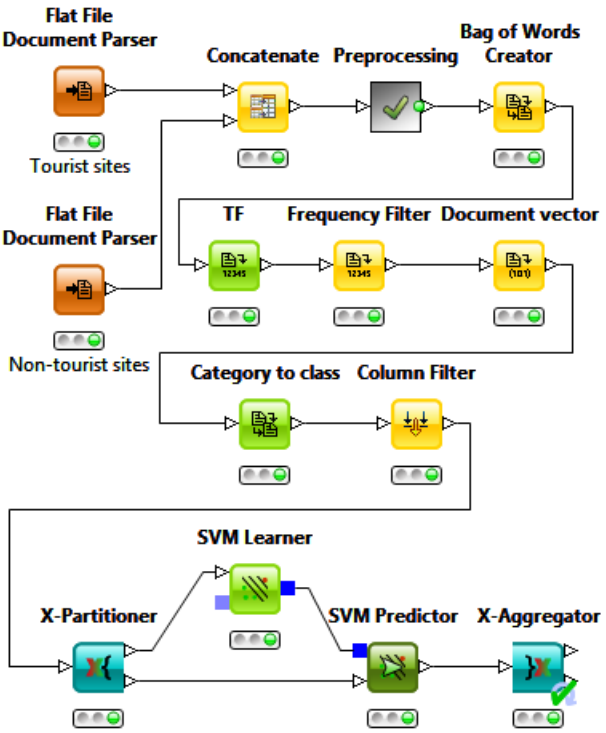


Figure 1. KNIME program used to validate training set.

| Row ID | Error in % |
|---|---|
| fold 0 | 5 |
| fold 1 | 15 |
| fold 2 | 10 |
| fold 3 | 2.5 |
| fold 4 | 10 |
| fold 5 | 7.5 |
| fold 6 | 10 |
| fold 7 | 15 |
| fold 8 | 15 |
| fold 9 | 12.5 |

Figure 2. Error rate table.

## E. *Preprocessing*

Any text written in any language contains a large number of elements common to all documents written in the same language, so these elements do not add any information content neither facilitate the document classification, rather they only increase computation time.

In the preprocessing phase, terms are manipulated in order to cut out elements that don't contain content, such as stop words, numbers, punctuation marks or to remove endings based on declination or conjugation by applying stemming.

While Puntuaction Erasure and Number Filter are the same for all languages, the other preprocessing block depend on the language to analyze, so we provided to Stop Word Filter a list of common terms in Italian language and used Snowball Stemmer, which allows the stemming of text in various languages (Fig. 3).

All terms survived to preprocessing are collected in a bag-of-word, which keeps the reference to the document belongs to, through this, the frequency with which a term appears in a document can be calculated and the terms less frequent can be removed. We set frequency filter threshold to 0,3%.

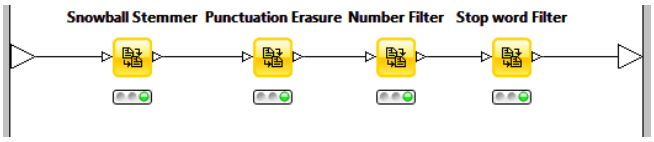


Figure 3. Content of preprocessing metanode.

## F. *Classification*

The classification consists of two phases: learning and prediction (Fig. 4).

The first phase is carried out by the SVM learner, i.e., a block that trains the SVM on the input data using a polynomial kernel:

$$K(x, y) = (gamma * x^T y + bias)^{power}$$

where:

- *x* and *y* are vectors in the input space;
- *power* is the degree of the polynomial;
- *bias* and *gamma* are hyperplane coefficients.

*Power* was the only parameter we set to 1 in order to have a linear SVM.

We have provided to SVM learner the dataset previously validated to get the best hyperplane for the separation of the two sets of classification: tourist text and non-tourist text.

In the second phase, the SVM predictor takes as input the model proposed by SVM learner and an unclassified dataset equal to 1/3 of the total number of text.

To prevent unclassified data were more than classified ones, each folder containing text supplied from sites of a given region has been divided into sub-folders in order to keep the aforementioned proportion and the classification procedure has been iterated several times till all the data has been classified.

 ISBN: 978-1-61208-510-4 

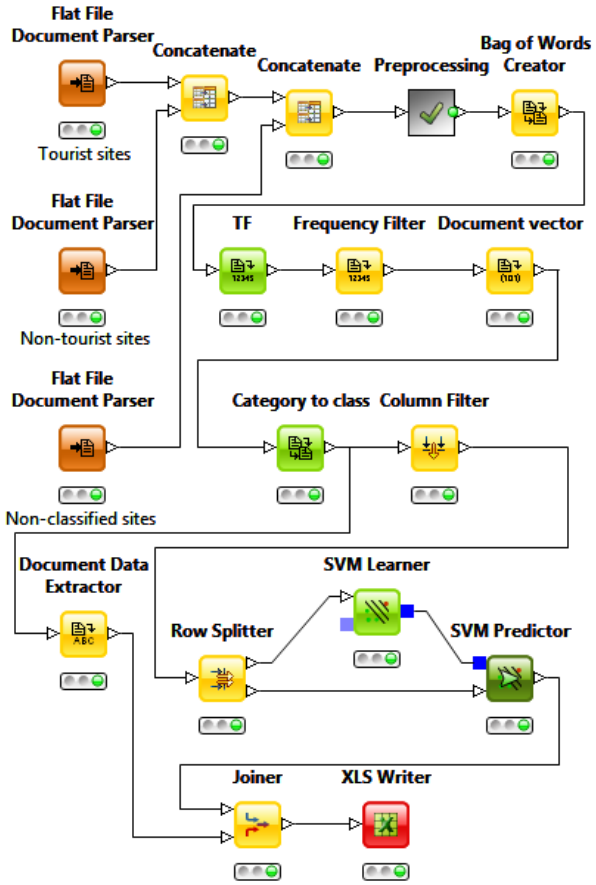


Figure 4. KNIME program used to classificate tourist and not tourist text.

## IV. PRELIMINARY RESULTS

Abruzzo has been the first analyzed region. The crawler, starting to seeds of certainly tourist sites, crawled 14.482 sites on a total of 2.377.378 pages. Using aforementioned classificator, we discover that only 4.200 sites were in Italian language and among these, only 2169 were classified as touristic.

Each of these sites was considered as a network node. In Fig. 5, we show the Abruzzo network after deleting non-Italian nodes. In Fig. 6, we show the Abruzzo (sub)network with tourism websites only.

To better characterize the subnetwork, we assess its in-degree (Fig. 7) and out-degree (Fig. 8) distributions in two graphs with logarithmic scale. Both graphs exhibit a trend that can be approximated to a power laws distribution, although further studies are needed to determine the exact degree and parameters.

Analyzing the network with standard bow-tie model (Fig. 9), the strongly connected core consists of 7 nodes only, whereas input nodes are only 3 and 85 are output nodes. Tendrils, i.e., pages that link to and from the In and Out group but are not part of either, are 1194. The disconnected component includes 880 nodes. Bow-tie graph shows that the network core, i.e., the strongly connected part, is really small and there are many disconnected nodes. Furthermore, the large number of tendrils highlights that most of the nodes prefer to be pointed, but rarely link other tourist sites.

The reason is probably that a low level of cooperation among tourism operators is present, and this negatively affects the development of Italian tourism.

TABLE I. ABRUZZO CRAWLING RESULTS

| Italian regional tourism authorities | #of sites crawled | # of pages crawled | #nodes | #of website in Italian language (#of nodes) | #of website classified as touristic (#of nodes) | #of connected components | #of nodes of the largest connected component |
|---|---|---|---|---|---|---|---|
| Abruzzo | 14482 | 2377378 | 14482 | 4200 | 2169 | 873 | 1289 |

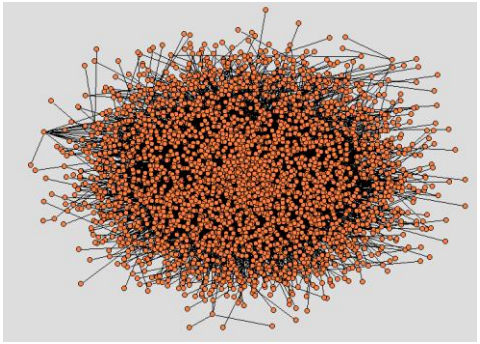

Figure 5. Italian nodes of Abruzzo network

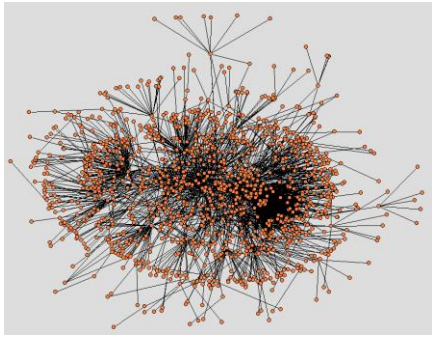

Figure 6. Tourism Italian nodes of Abruzzo network

 ISBN: 978-1-61208-510-4

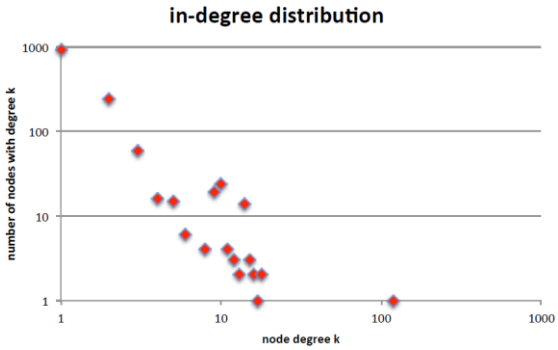


Figure 7. In-degree distribution of tourism nodes.

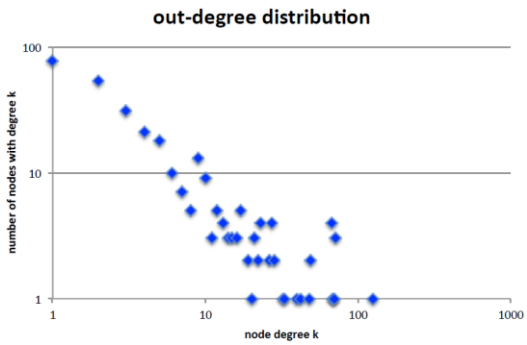


Figure 8. Out-degree distribution of tourism nodes.

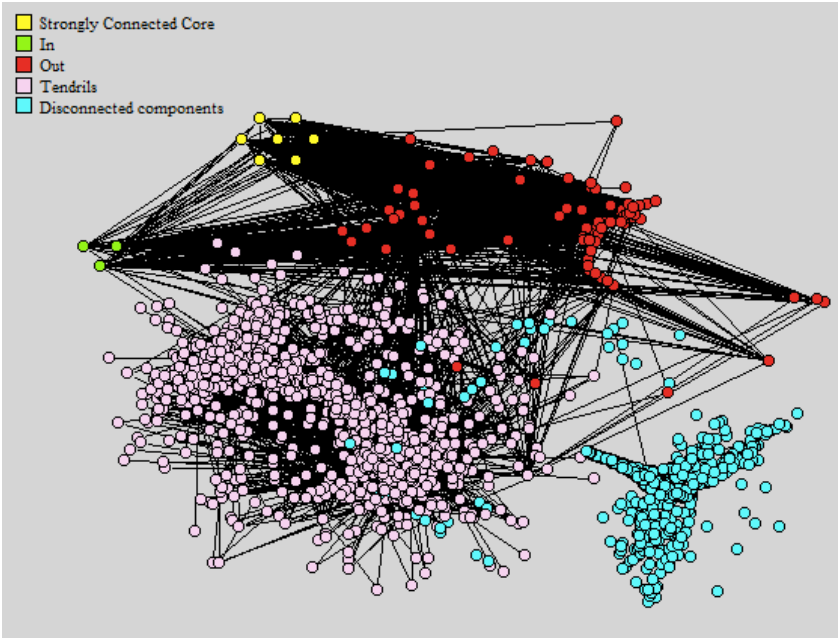


Figure 9. Bow-tie model of Abruzzo network.

## V. CONCLUSIONS AND FUTURE WORKS

In this paper, the Italian tourism websites network has been extracted from a large set of crawled sites. Using a proper seed of tourism websites and an SVM-based classifier, we crafted the tourism network, then we focused on one of the 20 italian regional networks, analyzing and briefly discussing its feature. In particular, a really low level of cooperation seems to characterize the network, and this is probably due to an extreme competition among tourist operators that negatively affects the management of tourism itself.

Although at a first stage, this work seems promising in terms of its dimension (a large website data set has been considered), becoming a significant base for further analysis. In particular we are going to craft all other regional Italian networks, in order to get a global view of the overall Italian tourism website network (actually, part of this work has been already carried out [19]). Finally, we are also planning to compare our approach with other classification models.

We believe that the further analysis of such global network will provide tourism stakeholders with significant and detailed data to improve tourism management, also by leveraging results coming from other research areas, for instance trust and ranking [20][21], to score and exploit "relevant" tourism sites, or recommendation systems [22], to discover and endorse "useful" tourism sites.

## REFERENCES

[1] J. Zhang, "Weighing and realizing the environmental, economic and social goals of tourism development using an analytic network process-goal programming approach." *Journal of Cleaner Production* 127: pp. 262-273, 2016.

[2] S. Pyke, H. Hartwell, A. Blake, A.Hemingway, "Exploring well-being as a tourism product resource."*Tourism Management* 55: pp. 94-105, 2016.

[3] L. Y. Y. Lu and J S. Liu, "A novel approach to identify research fronts of tourism literature." *Management of Engineering and Technology (PICMET), Portland International Conference on*. IEEE, pp. 2211-2217, 2015.

[4] A. Groth and D. Haslwanter, "Efficiency, effectiveness, and satisfaction of responsive mobile tourism websites: a mobile usability study."*Information Technology & Tourism*: pp. 1-28, 2016.

[5] I. Christensen, S. Schiaffino, and M. Armentano, "Social group recommendation in the tourism domain." *Journal of Intelligent Information Systems*: pp. 1-23, 2016.

[6] B. H. Farrell and L. Twining-Ward, "Reconceptualizing tourism."*Annals of tourism research* 31.2: pp. 274-295, 2004.

[7] F. Sebastiani, "Machine learning in automated text categorization." *ACM computing surveys (CSUR)* 34.1: pp. 1-47, 2002

[8] C. Lindemann and L. Lars, "Coarse-grained classification of web sites by their structural properties." *Proceedings of the 8th annual ACM international workshop on Web information and data management*. ACM, pp. 35-42, 2006.

[9] Z. Xu, I. King, and M. R. Lyu, "Web page classification with heterogeneous data fusion." *Proceedings of the 16th international conference on World Wide Web*. ACM, pp. 1171-1172, 2007.

[10] P. Calado, M. Cristo, E. Moura, N. Ziviani, B. Riberto-Neto, and M. A. Goncalves, "Combining link-based and content-based methods for web document classification." *Proceedings of the twelfth international conference on Information and knowledge management*. ACM, pp. 394-401, 2003.

[11] A. Sun, E. Lim, and W. Ng, "Web classification using support vector machine." *Proceedings of the 4th international workshop on Web information and data management*. ACM, pp. 96-99, 2002.

[12] http://developer.cybozu.co.jp/archives/oss/2010/10/language-detect.html [retrieved: July, 2016]

[13] A. Haugen, "The open graph protocol design decisions." *The Semantic Web–ISWC 2010*. Springer Berlin Heidelberg. pp. 338-338, 2010.

[14] A. Jovic, K. Brkic, and N. Bogunovic, "An overview of free software tools for general data mining." *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 37th International Convention on*. IEEE, pp. 1112-1117, 2014.

[15] M. R. Berthold et al., "KNIME: The Konstanz information miner:version 2-0 and beyond" . *AcM SIGKDD explorations Newsletter* 11.1: pp. 26-31, 2009.

[16] T. N. Phyu, "Survey of classification techniques in data mining."*Proceedings of the International MultiConference of Engineers and Computer Scientists*. Vol. 1., pp. 18-20, 2009.

[17] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques.": pp. 3-24, 2007.

[18] J. C. Platt, "12 fast training of support vector machines using sequential minimal optimization." Advances in kernel methods : pp. 185-208, 1999.

[19] G. Mangioni, A. Longheu, and R. Baggio, "The Italian Tourism Webspace: a Complex Network Analysis". Poster presented at the 5th Workshop on Complex Networks, Bologna, pp. 727-734, 2014.

[20] V. Carchiolo, A. Longheu, M. Malgeri, and G. Mangioni, "Trust assessment: a personalized, distributed, and secure approach" Journal Concurrency and Computation: Practice and Experience (CCPE) Special Issue: Special Issue on intelligent distributed computing, Volume 24, Issue 6, pp. 605–617, 2012

[21] V. Carchiolo, A. Longheu, M. Malgeri, and G. Mangioni, "Network size and topology impact on trust-based ranking", in press on Intl Journal of Bio-inspired Computation (IJBIC), http://www.inderscience.com/info/ingeneral/forthcoming.php?jcode=ijbic [retrieved: July, 2016]

[22] V. Carchiolo, A. Longheu, M. Malgeri, and G. Mangioni, "Context-based global expertise in recommendation systems" International Journal of Computing and informatics – Informatica - ISSN: 1854-3871 - Volume 34, no. 4, pp. 409-418, Slovenia, 2010.

 ISBN: 978-1-61208-510-4 