BrainMap – A Navigation Support System in a Tourism Case Study

Luís S. Teixeira, Rita A. Ribeiro, António Falcão, Gabriel P. Lopes, Ricardo Raminhos

BrainMap - A Navigation Support System in a Tourism Case Study

Luís F. S. Teixeira[1],Rita A. Ribeiro[2],António Falcão[2], Gabriel P. Lopes[1], Ricardo Raminhos[3],

[1]DI-FCT/UNL, 2829-516 Caparica, Portugal,
[2]CA3-Uninova, Campus – FCT/UNL, 2829-516 Caparica, Portugal,
[3]Viatecla - Estrada da Algazarra, 72, 2810-013 Almada, Portugal
[1]{lst, gpl}@fct.unl.pt
[2] {rar, ajf}@uninova.pt
[3]{rraminhos}@viatecla.com

Abstract. Presently, the amount of documents in corporations can make searching and browsing for a specific topic or information a very hard task. Therefore, it is important to develop tools to ease the retrieval of specific information and to support the exploration by users on corporate intranets (composed of several hundreds of gigabytes of documents). Although not explicitly identified, many of these documents are related among themselves (directly or implicitly). In this paper we discuss a navigation support system to explore graphs applied to document correlations, using a tourism case study.

Keywords: Document Correlation, Graph Exploration, Tf-Idf Metric, Jaccard Metric, Tourism

1 Introduction

The purpose of this paper is to discuss the capabilities of a Tourism navigation support system, developed in the scope of project BrainMap [1]. The motivation behind such a system is to provide a unique mechanism to find relations between unstructured documentation and provide the results in an innovative visual format that allows an intuitive navigation and data exploration. An illustrative case study, based on vacation packages of the tourism industry, is used to test the prototype.

Thus, this paper is structured as follows: Section 2 will briefly describe how this work and its components correlate with the conference theme. Section 3 will describe techniques and technologies used to achieve the objectives and some similar studies, as well as a sub-section describing the system design infrastructure. Section 4 will describe a usage example of the prototype. Section 5presents conclusions and future work.

2 Internet of Things

Considering the concept of the Internet of Things [2], and taking into account the idea that everything would be uniquely identified and connected, but not explicitly related, any mechanism that would allow the discovery and exploration of possible relationships would be very advantageous. In the context of this prototype, we can imagine that any document could have a unique identifier that would enable not only the identification of its content, but also other practical information that could be useful to extract relations to other documents with regard to their contents. A user would also be identified and have an associated profile containing preferences, usage domain, needs and expertise etc. That profile allows the system to better adjust the relations and gives the user the ability to interact with these documents in a more productive way. This relation and correlations would evolve in time with the interaction with documents, making the task of searching more focused, since the results would automatically have filtered themselves. These could be achieved by providing to the documents a unique identifier (usually termed a "URI" – Uniform Resource Identifier), which would further enhance the integration between physical objects and digital contents. The system presented in this paper would also allow navigation with seamless transition between physical objects and the digital contents.

3 Techniques and Technologies and Similar Studies

In this section, we describe the techniques and technologies used in the development of the navigation support system. We also present similar studies made in this scientific area. In the following sub-sections the document representation issue is reported, the extraction of key-terms is described and finally the correlation metrics problematic is described. Considering first similar studies, we have the work done in [3]where term and document correlation is addressed. Similarity metrics were used between the terms and the web documents. In [4] document correlation is established using a self-organizing map (SOM) to cluster documents by cluster of topics. There have been also attempts at visualizing the results retrieved by means of such systems. Considering the above references, one type of visualization is taken in consideration by [3] where the author projects in space a graph of correlated web documents while in [4] the authors represent the relationships between documents as a SOM (self-organizing map), to visually allow the user to observe the documents within the cluster and their neighbours. Our work mainly differs from the above, as we perform searches within a specific corporation intranet being solely based on textual document data, independently of the language or any other tools that would limit the support system. Specifically the authors of this work developed a navigation support [5], which is overviewed in the next sub-sections, and is then discussed in the scope of a customized tourism package case study.

3.1 Main Concepts of the Navigation Support System.

While developing this system, one of the first problems was the issue on document representation [5].There are many representations, specifically in the ontologies domain [6], but we followed a simple text-mining approach. Particularly, we use a "Bag of Words" [5] representation, i.e., the textual content is represented as a collection of all words in the text. To achieve this representation, a preprocessing step is necessary, which was accomplished by the use of scripts that removed unnecessary noise, e.g. html tags. The document text is indexed and therefore represented in a "Vector Space Model" [7] using Term Frequency – Inverse Document Frequency (Tf-Idf) [8] for term weighting purposes, as described below.

The first step in the system is to perform a request of information, using a word search query. The system will return a list of documents that contain the search keywords ordered by importance of the keyword. For accomplishing this task we used statistically based extraction methodologies. These are divided into two major types: approaches that use statistical methods and approaches that use alternative methods, which are not statistical or not purely statistical like the usage of grammars [9] or using statistical with linguistic modulation [10]. There are several ways to calculate metrics to give weight to the extracted words [5].As we are interested in the extraction, independent from the languages of the texts, and knowing that the statistical approaches are based mainly in frequencies counting, we have chosen to use a statistical approach, which uses a common term-frequency - inverse document frequency metric because it poorly scores the words that are not relevant, thus making unnecessary the explicit use of stop-lists for each of the languages envisaged[8].

Term Frequency – Inverse Document Frequency(Tf-Idf) [8] is a statistical metric used in information retrieval and text mining. It is used to evaluate how important a word is to a document in a corpus, increasing proportionally to the number of times a word appears in the document but it is offset by its frequency in the corpus. From [5] we use a probability, $p(W, d_j)$, in equation (1), defined in equation (2), instead of using the usual term frequency factor.

$$Tf - Idf(W, d_j) = p(W, d_j) * Idf\ (W, d_j) \tag{1}$$

$$p(W, d_j) = f(W, d_j)\ /\ Nd_j \tag{2}$$

$$Idf\ (W, d_j) = \log(\|D\| / \|\{d_j : W\ \in\ d_j\}\|) \tag{3}$$

Where $f(W, d_j)$ denotes the frequency of a word W in a document d_j and Nd_j stands for the number of words of d_j; $\|D\|$ is the number of documents of the corpus. So, $Tf - Idf(W, d_j)$ will give a measure of the importance of W within the particular documentd_j. By the structure of term Idf we can see that it privileges words occurring in fewer documents. The choice for using the Tf-Idf metric was

based on a comparative study performed by the authors of several metrics for the automatic extraction of document topics [5,11].

Correlation Metrics are needed to generate relationships between documents based on the score of their words. The terminology of networks and graph are used interchangeably, in this work we follow the same definition as presented in [12], where a network indicates a relationship between objects (free text documents in our case) and graph indicates relationships generated through an automatic process. According to [13] crisp relations, are relations of dichotomous type, like in binary logic, where a statement can be true or false and nothing else. Other important concept is the notion of similarity or proximity between elements of a corpus, for enabling us to learn how related they are and to construct the weighted graph. In this work we use Jaccard [14] similarity measure because it is flexible, simple to implement and is easily generalized to fuzzy weighted graphs, but any other similarity measure could have been used ([15][16]). Further, by using the minimum (min) and maximum (max) operators from the T-norms [13] the generalization of the Jaccard measure to the fuzzy interval of $[0,1]$ is done with the following formulation [17]:

$$J_{i,j} = \sum_{k} min(x_{ik}, x_{jk}) \Big/ \sum_{k} max(x_{ik}, x_{jk}) \tag{4}$$

Where the indexes i and j stand for the rows of the Tf-Idf Matrix (rows stand for documents, see Tf-Idf Matrix in Fig.).The x_{ik} means the column k (a Tf-Idf score of word) from the row i (a document). The definition for x_{jk} is similar but for line j (another document).

3.2 Navigation Support System Architecture

Having presented the background concepts and technologies required to understand the several parts of the navigation support system used in this work, the developed infrastructure is depicted in Fig.. It shows the flow and the four processing steps included in the navigation support system infrastructure

1. **Data preparation**: A company or corporation documentation has a large variety of formats, ranging from emails, project reports, studies reports, etc. In order to deal with this fact, normalization is required; this usually involves extracting the actual text from html pages (removing html tags). All required documentation is transformed automatically into plain text (txt) files. Several libraries are available (for many programming languages) that allow us to transform documents in a given format into a simple text file. For example [18] presents Apache Tika, a Java based toolkit for extracting content from a variety of document formats

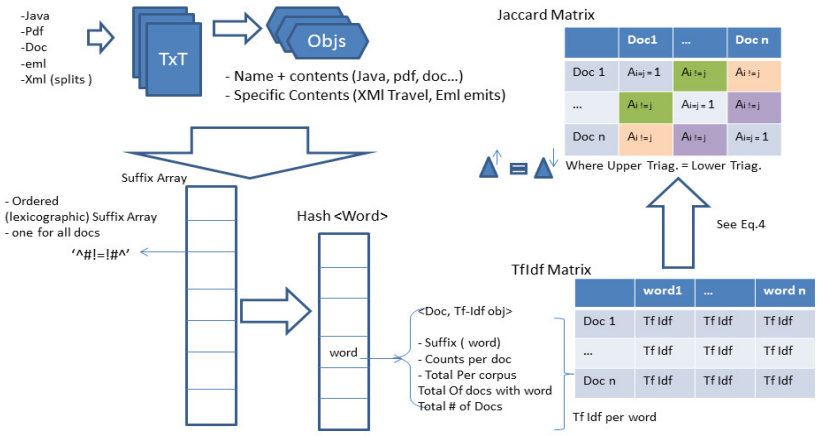

Fig. 1. Overall system design

3. **Calculate the weights of documents words**: For being able to mine any kind of relation between documents, their words (as we are using the bag of words representation) must be in some way weighed, assigning best scores to best words, i.e. best topic descriptors of the document content will have best scoring values. According to the work of [5,11] the best metrics to accomplish this are those based on Tf-Idf (Term Frequency – Inverse Document Frequency) and Phi-Square metrics.
4. **Mine Document Correlation**: This step goal is to discover document relations where they do not explicitly exist, just based on the documents textual content, requiring no semantic knowledge, and applying similarity metrics, as exemplified in the previous section example.

4 Tourism Case Study

The case study used is based on the Tourism industry as a source of documentation, in the form of textual descriptions from touristic destination packages [19]. These files are characterized by a small destination description, no more than a couple of sentences. Ideally, as more textual contents are available, the more accurate the Tf-Idf metric will be. Some other fields of the source file were also used, on one hand allowing more user interaction, and at the same time enriching the description initially offered. As an example, usually the country of the vacation package was not present in the description, but available in a specific field. This information is automatically appended to the description prior to the statistical treatment. After the system performs all the steps, the user has available a graphic interface application, allowing him to search for a destination, or for some characteristic that he prefers for his vacations, e.g. “beach” or “snow”, or “Bali”.

The results are presented to the user in a form of ranked list first (according to the score of the searched word in each vacation package).

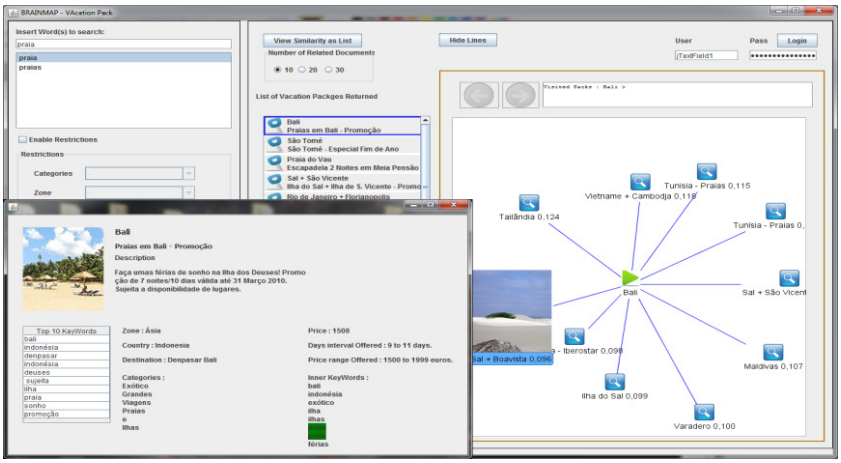

Fig. 1. User Interface overview

From this point the user can select one of the vacation packages presented, on which the application will then present him a graph with other vacation packages correlated with the user selection (see Fig. 1). At this point the user can consult the description of each vacation package; the interaction result is presented in Fig. 1. On this graph of packages the user may be presented with results that would not be seen on the ranked list, because it did not contain the exact search word, but will appear because the content of a description correlates to the initial selection of the user. This can be an interesting outcome, while searching for a vacation destination.

The similarity distance given the Jaccard metric value, can be visually represented as the distance between the center node and the related nodes, see Fig. 2

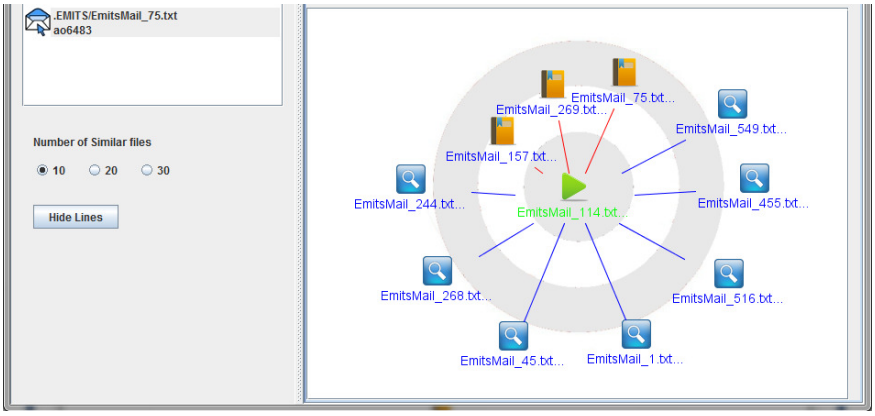

Fig. 2. Correlation graph with Jaccard distances represented by arc lengths.

See the illustrative example below. The system would work for a corpus of two documents composed by a bag of 4 words that would generate a matrix containing the Tf-Idf values of the words in their documents.

Table 1. The TfIdf Matrix for the example documents.

Docs\words	praia (beach)	ilhas (islands)	exótico (exotic)	tropical
Thailand	0.7	0.3	0.5	0.4
Bali	0.6	0.8	0.4	0.6

Now we want to find the Jaccard Value (using equation 5) between Thailand (Doc 1) and Bali (Doc 2) in Table 1. To calculate the correlation between lines 1 and 2, we start by setting i =1 and j = 2 in equation (5), determining the maximum and minimum for each word. For instance, in column 1 ("beach") the maximum of (0.7, 0.6) is 0.7 and the minimum is 0.6. We do this for all words and then using equation 4, we calculate the Jaccard measure:

$$J_{i=1j=2} = \frac{0.6 + 0.3 + 0.4 + 0.4}{0.7 + 0.8 + 0.5 + 0.6} = 0.65 \quad (5)$$

The result of 0.65 depicts the correlation between Thailand and Bali using the Jaccard metric. This result means that Thailand is similar to Bali in 0.65. Notice that 0 means completely dissimilar and 1 means total similarity (this result only happens when correlating the same document with itself).

5 Conclusions and Future Work

This paper described an application of tourism industry using a novel Navigation Support System. This application demonstrated how the system supports intranet corporate users in the search for either/or keywords and documents, and also enable users to navigate across a network of related documents. Visual inspection of the correlated documents returned by the system, showed similar contents that would allow a user to navigate through a graph and possibly allowing him to reach a document that otherwise would not be found, e.g. does not contain the initial search word. We believe this type of application can be very useful for exploring and finding knowledge and information within the massive databases of corporations. Moreover, it can help novices to learn about the "business" in a faster and user-friendlier way.

As future work we will consider evolving the prototype to the web environment, making it readily accessible from multiple platforms. Other improvements such as considering as keywords of a document not only words but also multi-words, as indicated in the work of [5] will also be considered. Additionally in translations tasks, it is possible to enhance the translation process by finding correlations between already translated documents that are not necessarily parallel to the current document (i.e. not translations of each other).

Acknowledgments he work on this paper was developed in the context of the project "BrainMap" lead by the company ViaTecla (Portugal) in collaboration with UNINOVA and University of Évora (both in Portugal). Financedbythe Portuguese "QREN- Quadro de Referência Estratégico Nacional; Programa Operacional de Lisboa". This was also supported by the Portuguese Foundation for Science and Technology (FCT/MCTES) through funded research projects ISTRION (ref. PTDC/EIA-EIA/114521/2009) and VIP-ACCESS (ref. PTDC/PLP/71142/2006).

References

1. ViaTecla (2011) BrainMap. http://www.viatecla.pt/inovacao/brain_map.
2. R. Van Kranenburg, E. Anzelmo, A. Bassi, D. Caprio, S. Dodson, M. Ratto (2007) The Internet of Things, vol 2. A critique of ambient technology and the all-seeing network of RFID, Network Notebooks.
3. Viji S (2002) Term and Document Correlation and Visualization for a set of Documents. Stanford University,
4. Klose A, Nürnberger A, Kruse R, Hartmann G, Richards M (2000) Interactive text retrieval based on document similarities. Physics and Chemistry of the Earth, Part A: Solid Earth and Geodesy 25 (8):649-654
5. Luís F. S. Teixeira, Gabriel P. Lopes, Rita A. Ribeiro (2012) An Extensive Comparison of Metrics for automatic extraction of Key Terms. Proceedings of the 4th International Conference on Agents and Artificial Intelligence (ICAART 2012), Algarve, Portugal.
6. Maynard D, Peters W, Li Y (2006) Metrics for Evaluation of Ontology-based Information Extraction.WWW2006, Edinburgh, UK, May 22–26
7. G. Salton, A. K. C. Wong, C. S. Yang (1975) A vector space model for automatic indexing. Commun ACM 18 (11):613-620. doi:10.1145/361219.361220
8. Gerard Salton, Chris Buckley (1987) Term Weighting Approaches in Automatic Text Retrieval. Cornell University.
9. Taniza Afrin (2001) Extraction of Basic Noun Phrases from Natural Language Using Statistical Context-Free Grammar. Virginia Polytechnic Institute and State University,
10. Anette Hulth (2003) Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In: EMNLP '03 Proceedings of the Conference on Empirical Methods in Natural Language Processing. ACL, Stroudsburg, PA, USA, pp 216 - 223
11. Luís F. S. Teixeira, Gabriel P. Lopes, Rita A. Ribeiro (2011) Automatic Extraction of Document Topics. In: Camarinha-Matos LM (ed) DoCEIS'11 - 2nd Edition of the Doctoral Conference on Computing, Electrical and Industrial Systems, vol 349. Technological Innovation for Sustainability. IFIP International Federation for Information Processing, Caparica, Portugal, pp 101–108. doi:10.1007/978-3-642-19170-1
12. Rada Mihalcea, Dragomir Radev (2011) Graph-based Natural Language Processing and Information Retrieval. . Cambridge University Press, New York, NY 10013-2473
13. Hans-Jürgen Zimmermann (2001) Fuzzy Set Theory and its Applications. 4th ed. Springer,
14. S. Miyamoto (1990) Fuzzy Sets in Informational Retrieval and Cluster Analysis, vol 4. Series D: System Theory, Knowledge Engineering and Problem Solving, 1st ed. Springer,
15. C. Shyi-Ming, Y. Ming-Shiow, H. Pei-Yung (1995) A comparison of similarity measures of fuzzy values. Fuzzy Sets Syst 72 (1):79-89. doi:10.1016/0165-0114(94)00284-e
16. Amit S, Chris B, Mandar M (1996) Pivoted document length normalization. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, Zurich, Switzerland.
17. Luis M. Rocha, Tiago Simas, Andreas Rechts, Mariella Di Giacomo, Rick E. Luce (2005) MyLibrary at LANL: proximity and semi-metric networks for a collaborative and recommender Web service. In: In The 2005 IEEE/WIC/ACM International Conference on Web Intelligence. IIEEE Press, pp 565-571. doi:10.1109/WI.2005.106
18. Chris A. Mattmann, Jukka L. Zitting (20 12) Tika in Action. Manning Publications Co.
19. Vacation packages database (2011) Viatecla. www.viatecla.com.