Content Based Visual Mining of Document Collections Using Ontologies

Katia Romero Felizardo[1], Rafael Messias Martins[1], José Carlos Maldonado[1], Alneu de Andrade Lopes[1], and Rosane Minghim[1]

Institute of Mathematics and Computer Science, University of Sao Paulo, Sao Carlos, SP, Brazil
{katiarf, rafaelmm, jcmaldon, alneu, minghim}@icmc.usp.br

Abstract. Document collections are important data sets in many applications. It has been shown that content based visual mappings of documents can be done effectively through projection and point placement strategies. An important step in this process is the creation of a vector space model, in which terms selected from the text and weighted are used as attributes for the vector space. That step in many cases impairs the quality of the projection due to the existence, in the data set, of many terms that are frequent but do not represent important concepts in the user's particular context. This paper proposes and evaluates the use of ontologies for content based visual analysis of textual data sets as a means to improve the displays for the analysis of the collection. The results show that when the ontology effectively represents the data domain it increases quality of maps.

1 Introduction

The increase of the quantity of information made available every year in digital form is notorious. However, extracting value from this set of information has also become progressively more difficult [3].

An approach to overcome this information overload is to use Text Mining for automated extraction of patterns and models from collection of documents. Applying Text Mining techniques involves a pre-processing stage, responsible for loading, integrating, cleansing, structuring and normalizing data, usually in a representation referred to as Vector Space Model - VSM [8]. To build a VSM representation, each document is transformed into a feature vector in a multidimensional space. In this vector, features are terms considered relevant found in the document collection, and coordinates are some sort of weighted term count. Since this model is limited to the terms explicitly selected from the texts, some approaches for extending it were proposed, including ontology-based ones [2,10]. These new proposals mean to include not only explicit information on the document vectors, but also semantic relationships between the terms in the collection of documents.

An approach to Text Mining includes Visualization techniques, configuring a Visual Text Mining scenario [4]. In this context, exploring a collection of documents involves an iterative and interactive process over a graphical representation of that collection. A software tool that provides an opensource visualization environment for this process is the Projection Explorer (PEx) [7]. PEx uses a vector space model to structure, compare

and calculate the distances between documents, in order to group and project them onto a two-dimensional space in such a way that similar documents are placed closely in the final display, according to the chosen distance measure.

In this paper we evaluate the extension of the vector space model, based on the use of ontology, for Visual Mining purposes. For that, PEx was adapted with a new vector space construction engine, which considers concepts and synonyms instead of only terms. One case study was performed to visually compare different projections of the same document collection – with and without the use of domain-specific ontologies.

The remaining of this paper is organized as follows: Section 2 presents the vector space model, some of the proposals for its extension – specially ontology-based ones – and an overview of the PEx tool. The case study performed is detailed in Section 3, which also presents the evaluation of the results; final considerations are presented in Section 4.

2 Related Work

Currently, visual exploration of document collections is a topic of interest to the scientific community. Two of the challenges of visualizing document collections are the lack of an explicit 2D or 3D representation of the documents and the high dimensionality of the data [3]. In VSM [8], also known as "bag of words", each document is represented by a vector of *n* dimensions, where *n* represents the number of different terms found on the collection.

Since the model considers only terms explicitly found in the documents, it is limited, since human writing is characterized by an extensive use of synonyms and a good number of terms is not relevant within a particular context. The probability that two researchers use the same term to refer to the same concept is often lower than 20% [9]. Thus, a direct comparison of terms may not be sufficient. Because of this, new approaches, based on ontologies, have been proposed.

According to Spasic et al. [9] the task of connecting textual information with an ontology is arduous, but this connection can be reached through terms; in other words, it's the terms found in the texts that map the specific domain concepts, represented in the ontology. In this context, the work of Yoo and Hu [10] describes the construction of the vector space based on mapping terms into concepts of an ontology. The proposed process begins by the conversion of documents into an adequate format, reducing the number of considered terms, removing *stopwords*, and selecting only terms between 1 and 3 – *grams* (sentences composed by 1, 2 or 3 words) as term candidates.

Then, to incorporate the knowledge contained in the ontology, candidate terms of the documents are mapped into concepts of the ontology. By doing this, terms are replaced by concept descriptors which unifies synonyms and related terms. For example: with the use of the Medical Subject Headings (MeSH) ontology, terms like "Cancers", "Tumors" and "Benign Neoplasms" are all identified by the "Neoplasms" concept descriptor [10].

In the work by Hotho, et al. [2], different strategies are investigated for using ontologies in the construction of vector spaces. As with the previous work, concepts are identified by entry terms, which can be the concept itself or its synonyms. The three

proposed strategies are: (i) **Add Concepts ("add"):** Each vector is formed by adding to the terms found in the document the corresponding concepts. If a term found in the text corresponds to an ontology concept, the concept is added to the vector (and so is the term); (ii) **Replace Terms by Concepts ("repl"):** The second strategy consists on replacing terms by their corresponding concepts (when they exist). Terms that have no corresponding concept on the ontology are still considered on the document vector; and **(iii) Concept Vector Only ("only"):** The third strategy is similar to the second, but in this case, terms that have no corresponding concept on the ontology are discarded. The resulting vectors are composed only of concepts from the domain of the ontology.

3 Visual Analysis of Ontology Based Projections

The case study presented in this section aims at evaluating the use of ontologies for vector space construction in the context of Visual Text Mining. Used concepts and strategies are based mainly on the works of Spasic, et al. [9], Hotho, et al. [2] and Yoo and Hu [10], presented in the previous section. We chose a corpus on Software Testing, which has an already established ontology called OntoTest [1].

PEx, the software tool extended for this case study, is a generic platform for visual mining and exploration of document collections, using the vector space model. Based on the vector representation, distance between document are computed, resulting on a distance matrix with high dimensionality – which makes it hard to accomplish good 2D representations of the information and may impair interpretation. As an alternative, dimensionality reduction techniques (projections) – e.g. Least Square Projection (LSP) and Interactive Document Map (IDMAP) [6] – are used, allowing multidimensional data to be displayed in a 2D space. Such projections are defined based on different criteria, commonly trying to preserve distance relations between points on their original multidimensional space.

In the extended PEx, the use of an ontology in the text mining process is initiated by loading an XML file containing the ontology. It contains a set of concepts related to a specific domain, where each concept has synonyms that allow its identification in the text. The main role of the ontology is to distinguish terms that represent relevant domain concepts from terms that are considered irrelevant and should be discarded.

If an ontology is defined, a concept is identified in the text by its own presence and also by its synonyms. Thus, if a synonym of a concept present in the bag of words is found in the text, then the concept frequency will be updated including the frequencies of its synonyms. The result is a collection of [Concept, Frequency] pairs, which is then transformed into a weighted attribute-value vector. If an ontology was not chosen, the vector space is built as usual, with the document n-grams (words or phrases) extracted from the text files and their frequencies (weights) calculate, after eliminating stopwords and applying stemming.

It is important to notice that in both cases (with or without the ontology), the final product of this stage is the same – a distance matrix calculated from the corpus' vector space representation – so the rest of the tool's operation remains the same. To evaluate how the use of an ontology on the vector space construction affects visual mappings, several document maps were generated – with the new version of PEx – and compared,

both with the common and the ontology-based vector space approaches, as reported in the next Section.

3.1 Case Study – Software Testing Domain

The corpus for the case study was composed of 118 articles, originated from five systematic reviews. The topic of the systematic reviews were the following: Group 1 – Integration Testing of Aspect-Oriented Programs; Group 2 – Verification, Validation and Testing of Web-Services Composition; Group 3 – Mutation Testing of Aspect-Oriented Programs; Group 4 – Software Testing on Agile Methods; and Group 5 – Quality of Component-Based Software Development Processes. Each of these subsets of papers was considered as a pre-classification of the corpus, and was represented in the projections by a different color, as follows: Group 1) dark blue, Group 2) yellow, Group 3) Green, Group 4) Red and Group 5) light blue.

The IDMAP technique and cosine as distance measure was used for projecting the data points. To map documents' terms into domain concepts, an ontology of software testing – called Ontotest [1] – was used. Different perspectives involved in the testing activity, such as techniques and criteria, human and organizational resources, and automated tools are explored in this ontology.

Visual analysis was adopted to evaluate the results, using a technique of coordination by identity between different visualizations. This technique highlights the same individuals in all current visualizations when they are selected in one of them. For example, one can see in Figure 1, a coordination between (a) and (b): elements of the group $G1$, selected in (a), are highlighted in (b). For each set of items two projections were created, one with and one without the use of the ontology.

Evaluation of Case Study Projections: The coordinated view of the documents contained in the set about "Integration Testing of Aspect-Oriented Programs" are represented in Figures 1 (a), (b), (c) and (d) (G1 or dark-blue points). Comparing the positions of documents in the alternate projections, it's possible to notice that in ontology-based projections – on the left (Figures 1 (a) and (c)) – although the documents are not tightly grouped (placed closely), there is an indication of two small groups and a decrease in scattered points. On the other hand, in the projections without the ontology – on the right (Figures 1 (b) and (d)) – the documents are all spread over the map.

The group of papers about "Verification, Validation and Testing of Web-Services Composition" (G2 or yellow points) presented the best result, with the points being placed very closely in the ontology-based projection in Figure 1 (e). In contrast, in the projection without ontology – Figure 1 (f) – the documents are scattered in the center of the map.

The coordination that refers to the set of documents on "Mutation Testing of Aspect-Oriented Programs" (G3 or green points) is shown in Figures 1 (g) and (h). Again, comparing the positioning of documents in the two projections, we conclude that in the ontology-based projection – Figure 1 (g) – even though the documents are not completely grouped, there is a lower dispersion. On the other hand, in the projection without ontology – Figure 1 (h) – the documents are scattered throughout the whole eastern region of the map.

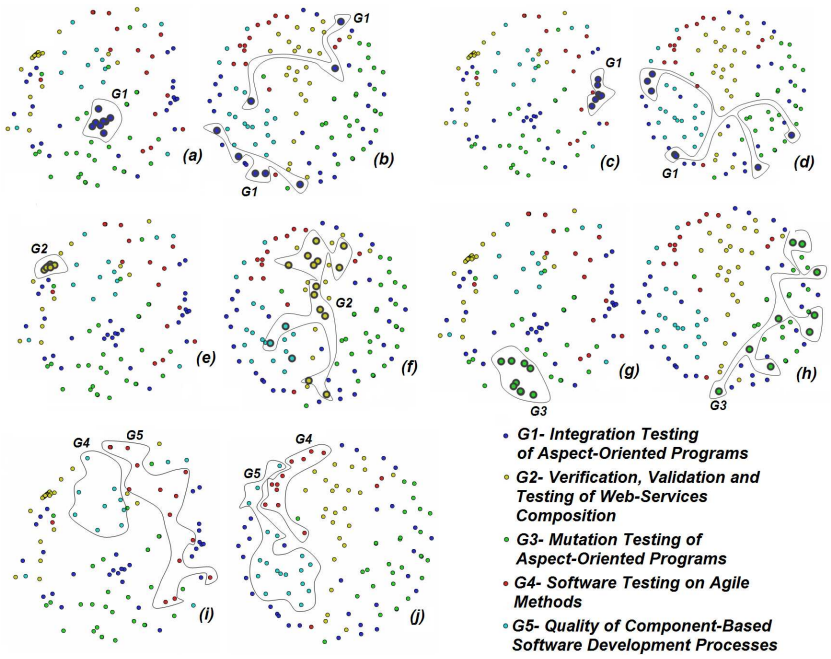

Fig. 1. Coordinations between projections: with (left) and without (right) the ontology

The other sets of documents, “Software Testing on Agile Methods” (G4 or red points) and “Quality of Component-Based Software Development Processes” (G5 or light-blue points), showed no significant positioning between the projections (Figures 1 (i) and (j)), with documents appearing scattered in both cases. It is worth noticing that OntoTest does not have concepts related specifically to the research themes involved in these reviews.

Even considering the good results of the projections presented, some points should be noticed: (i) the rate of development of any research area can turn an ontology outdated or limited quickly, failing in the task of representing relevant information from more recent documents. Spasic, et al. [9] also warned that ontologies become incomplete as a result of rapid expansion of knowledge, and this is one of the obstacles in their use on text mining; and (ii) the absence of some concepts makes document representation difficult and, as consequence, affects the final projection. The documents’ authors do not necessarily follow writing conventions of the ontology, so it is possible that there are key terms in the texts that are not identified. Therefore, it is important for the ontology to be flexible and to accept changes in the writing of the same concept.

An improved version of this case study is presented on the following section, based on the conclusions obtained. Some new concepts and synonyms were included in the ontology to try to consider more of the new concepts expressed in the corpus.

3.2 Improvements on the Case Study

The results presented in the previous section showed that the subset of documents about "Software Testing on Agile Methods" were among the worst results when using an ontology to support the projection. One of the possible causes for this is that the ontology did not include concepts from the agile development domain; documents of this area were not well represented by the ontology-based vector space model. In order to confirm this hypothesis, some concepts specifically related to the agile testing domain were identified and Ontotest was updated to include them. According to the work of Melnik [5], the concept "Acceptance Tests" is used by many different authors with many different terms, for example: "functional tests", "customer tests", among others. This concept and all its synonyms were added to Ontotest, along with the concepts "Test Driven Development" and "Agile Software Testing".

To improve the visibility of the new results, 114 new documents about "Software Testing on Agile Methods" were added to the corpus, which is now composed of a total of 232 documents. Based on the new total number of documents, the LSP technique and cosine as distance measure was chosen for this new version of the case study, since it generally presents better results than IDMAP for larger data sets. The rest of the setup for the projections was the same from the previous version.

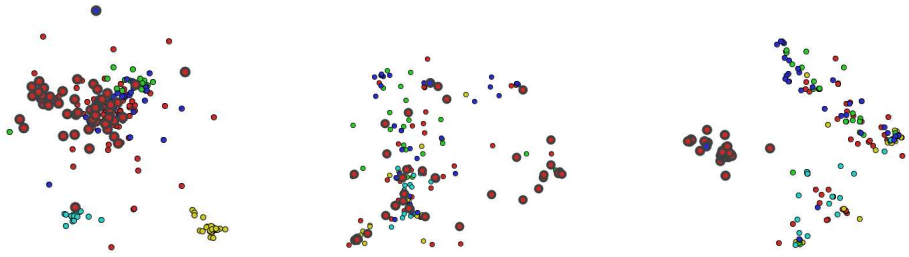

Fig. 2. New projection results: (a) without Ontotest, (b) with Ontotest without agile testing concepts, and (c) after updating Ontotest with agile testing concepts

Results using the improved version of the ontology are shown on Figure 2, with the coordinated view of the agile testing documents (highlighted red points in the three images). The first two projections (Figures 2(a) and 2(b)) repeat the results from the previous execution of the case study, using no ontology at all and the original ontology (without agile testing concepts), respectively. On the other hand, the projection on Figure 2(c) – which uses the updated Ontotest – shows two main improvements over the two previous projections: the agile testing documents became densely grouped, unlike Figure 2(b); and they are separated from the rest of the corpus, unlike Figure 2(a).

It is important to notice that only three specific concepts – and a few synonyms – were added to the ontology, and it was enough to improve substantially the quality of the projections for this domain. These satisfactory results motivated a new ontology update and a new execution of the case study, this time by adding concepts related

to three other subsets of documents: Group 2 – “Verification, Validation and Testing of Web-Services Composition” (5 concepts added), Group 3 – “Mutation Testing of Aspect-Oriented Programs” (4 concepts added) and Group 5 – “Quality of Component-Based Software Development Processes” (20 concepts added). The resulting projection is shown in Figure 3.

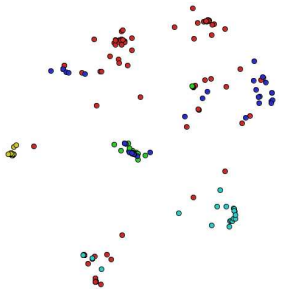

Fig. 3. New projection result after updating Ontotest for the second time

The improvements in grouping and the separation of groups 2 (yellow) and 5 (light blue) are visibly good, specially when compared to previous results (Fig. 2). Groups 1 and 3 got mixed together in the middle of the projection, probably because their themes are similar – testing techniques for aspect-oriented programs.

4 Conclusions and Future Works

This article presents a proposal for the use of ontologies instead of corpus-extracted vocabulary in the visual analysis of document collections, considering not only the terms explicitly found in the document but also information related to the context and the domain.

To implement the idea, the PEx tool was adapted and a XML file format was defined for the storage of ontologies, with support for concepts and synonyms. The evaluation was conducted on a case study using a corpus from the Software Testing corpus, which had an already established ontology (OntoTest) and a pre-classification done by researchers.

In general, one can see that the idea of using ontologies to improve the visualization of documents sets is promising. However if the information domain is not effectively represented, with all the possible variations of a concept, the use of an ontology can impair the resulting map. Instead, if the field is effectively represented resulting maps increase its quality.

Despite the improvements achieved, some problems and limitations should be considered when one wishes to use ontologies to help the view. The speed with which research areas develop may lead to an ontology becoming quickly outdated, failing in the task of representing relevant information from more recent documents.

As future work, similar case studies are planned to analyze other ontology-based vector space construction strategies, specially “repl” and “add”, by Hotho, et al. [2]. Other types of knowledge represented in the ontologies – like hierarchical relationships between concepts – are also being investigated in the context of visual mining.

References

1. E. Barbosa, E. Nakagawa, A. Riekstin, and J. Maldonado. Ontotest: An ontology of software testing. In *Workshop on Ontologies and Metamodeling in Software and Data Engineering (WOMSDE 2007)*, 2007.
2. A. Hotho, S. Staab, and G. Stumme. Ontologies improve text document clustering. *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 541–544, Nov. 2003.
3. D. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.
4. A. A. Lopes, R. Pinho, F. Paulovich, and R. Minghim. Visual text mining using association rules. *Comput. Graph.*, 31(3):316–326, 2007.
5. G. Melnik, F. Maurer, and M. Chiasson. Executable acceptance tests for communicating business requirements: customer perspective. pages 12 pp.–46, July 2006.
6. F. V. Paulovich and R. Minghim. Hipp: A novel hierarchical point placement strategy and its application to the exploration of document collections. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1229–1236, 2008.
7. F. V. Paulovich, M. C. F. Oliveira, and R. Minghim. The projection explorer: A flexible tool for projection-based multidimensional visualization. In *SIBGRAPI ’07: Proceedings of the XX Brazilian Symposium on Computer Graphics and Image Processing*, pages 27–36, Washington, DC, USA, 2007. IEEE Computer Society.
8. G. Salton, A. Wong, and C. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.
9. I. Spasic, S. Ananiadou, J. McNaught, and A. Kumar. Text mining and ontologies in biomedicine: making sense of raw text. *Brief Bioinform*, 6(3):239–251, September 2005.
10. I. Yoo and X. Hu. Biomedical ontology mesh improves document clustering qualify on medline articles: A comparison study. In *CBMS ’06: Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems*, pages 577–582, Washington, DC, USA, 2006. IEEE Computer Society.