Tight Coupling of Personal Interests with Multi-dimensional Visualization for Exploration and Analysis of Text Collections

VinhTuan Thai, Siegfried Handschuh, Stefan Decker
Digital Enterprise Research Institute (DERI), National University of Ireland, Galway
firstname.lastname@deri.org

Abstract

In this paper, we present an interactive matrix-based multi-dimensional visualization component which enables the users to explore a text collection along different conceptual dimensions. Of importance in our approach are the tight coupling of the users' personal ontologies representing their spheres of interest with the visualization component and the application of barycenter heuristic for edge crossing minimization to enhance its visual display. We also discuss how IVEA, the information visualization tool containing the proposed component, can address the commonly perceived constraints of building a personal ontology from scratch for IVEA to work.

1 Introduction

Information visualization is a key mechanism to gaining insights from text collections and is thus useful for a wide spectrum of users, from scientists and analysts to news readers. The insights obtained from the exploration and analysis of text collections can enable the users to understand the distribution of topics or to identify trends and linkages between different entities [18]. While many tools support document corpus visualization (e.g. [18, 24, 4]), most present findings that are independent of the users' interests. Their focus is usually to identify the main entities (e.g. topics, people, locations) within text collections and then visualize different linkages between them. While the automatic extraction of entities is helpful, it is also important that the visual exploration process can be aligned with the users' personal interests, especially when certain entities within their spheres of interests are of significant importance to their exploration goals. As a result, with the existing tools, the users cannot have a personal lens via which they can focus on some particular entities and relationships while exploring text collections.

To address the above issue, we proposed an innovative document-centric approach toward exploratory visualization of a text collection by:

- Involving the users at an early stage in the visualization process, whereby they define their spheres of interest by encoding the important entities and their relationships into a personal ontology.
- Leveraging upon the hierarchical structure of entities defined in the abovementioned ontology to allow the users to interpret various aspects of a text collection at different levels of detail.
- Employing coordinated and multiple views to allow the users to look at documents in a text collection from different perspectives.
- Suggesting the users with frequent phrases within documents to keep the personal ontology updated with new entities potentially matching their evolving interests on-the-fly. With the newly added entities, the personal ontology becomes a richer and better representation of the users' interests and hence can lead to more personalized subsequent exploration and analysis experiences.

Details on the initial design and implementation (a tool called IVEA) of the proposed approach can be found in [23]. Based upon the visual information-seeking mantra: "*Overview first, zoom and filter, then details-on-demand*" [19], the initial design of IVEA consists of four coordinated views:

- A *personal knowledge view* which shows the hierarchical structure of concepts and instances within the personal ontology and serves as an anchor for the exploration process
- An *overview display* via a scatter plot which shows the relevance of documents with regard to the concepts or instances placed on its two dimensions

- A group of *detailed views* using bar charts which displays the relevance values and frequencies of each instance appearing in the document being examined as well as the frequent phrases which can be used to enrich the personal ontology

- An *entities distribution view* which employs a variant of the TileBars [6, 16] to display the relative locations of appearance of the ontology's instances together with their corresponding frequencies in each part (fragment) of a document [23].

While the initial user feedback on IVEA was positive, it also highlighted a shortcoming in the overview display. Although the scatter plot allows the users to focus on documents which are relevant to the concepts/instances placed on its two axes, its inherent limitation is that the users can only have an overview of how relevant the documents in a collection are with respect to *two* entities of interest at a time. Initial user feedback indicates that the ability to explore a text collection along multiple conceptual dimensions at the same time is essential. In addition, a suitable visualization component also needs to allow the users to dynamically explore at different levels of granularity based on the hierarchical structure between a class and its subclasses or instances, as defined in the personal ontology.

To meet the above requirement, we propose in this paper an interactive matrix-based multi-dimensional visualization component which is tightly coupled with the users' personal ontology. In addition, we also apply the barycenter heuristic for one-sided edge crossing minimization [22, 12] to enhance the display of this component.

The remainder of this paper is organized as follows: In Section 2 we highlight related work. The proposed solution is described in Section 3, followed by a discussion in Section 4. Finally, we conclude the paper and outline future work in Section 5.

2 Related Work

A good overview of multi-dimensional visualization approaches can be found in [10]. Of importance are fundamental ones such as: parallel coordinates [8], relational table [21, 15], iconic displays [6], circle segments [1], star coordinates [9] and dimensional stacking [11]. Apart from these useful generic techniques, a number of tools are designed specifically for providing an overview of text collections over various dimensions. DocCube [13] is based on OLAP principles and employs 2D and 3D scatter plots to show the number of documents belonging to several categories, which are grouped into dimensions. DocCube, however, does not allow for relative comparison between documents to see how they are relevant to a set of concepts or categories. It is also limited to displaying at most 3 dimensions at a time. FeatureLens [2] is also an information visualization tool for exploring text collections. However, it focuses on presenting the relative locations and supports of text patterns mined from text collections. While the FeatureLens's interface is very intuitive, it does not allow for visual exploration at different levels of detail based on a hierarchical structure of concepts. Other work such as [18, 24, 4] provide no mechanism for the users to incorporate a personal ontology into the visual exploration process. It is also worth noting AutoFocus, an ontology-based visualization tool which displays search results for documents on the desktop as clusters of populated concepts [3]. While allowing for the exploration of a document collection based on the extracted metadata and search keywords, no deep insights or comparisons can be attained with AutoFocus when the degrees of relevance of documents with respect to each entity within the users' spheres of interests are not available. To the best of our knowledge, we are not aware of any existing work which tightly couples a personal ontology defining a user's sphere of interest with a multi-dimensional visualization component to enable visual exploration of a text collection.

3 Proposed Solution

In this section, we first present the background information on how the relevance of each document in a text collection with respect to an instance in the personal ontology is measured. Then, different visual and interactive aspects of the multi-dimensional visualization component are described in detail.

3.1 Background

In order to obtain information on how relevant each document in a text collection is with respect to the concepts and instances representing the entities of interest to the users, we extract the texts of all documents and store them into a Lucene[1] index. A Boolean query (with the default OR operator) consisting of all instances in the personal ontology is used to retrieve documents which contain terms matching the users' interest. The term weight in a document of an instance is used as the relative relevance value of that document with respect to that instance. In Lucene, a variant of the well-known TF.IDF term weight function is used, which takes into account the frequency of a term locally (in a document) and globally (in the whole collection), as well as the length of a document itself [5]. The relative relevance value of a document with respect to a class is the aggregated relevance value of that document with respect to all of its direct instances and recursively, all of its subclasses.

[1] http://lucene.apache.org/

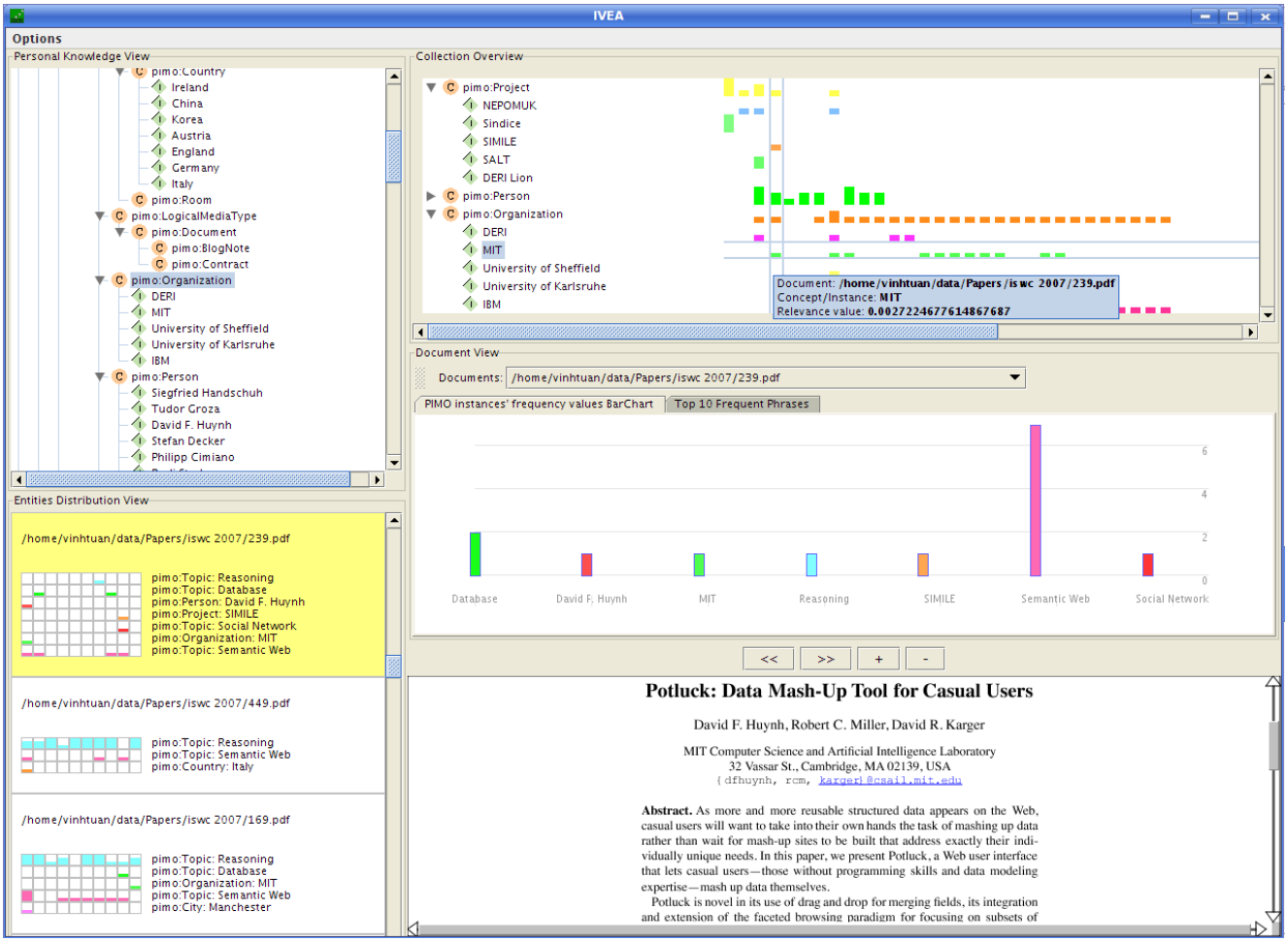

Figure 1. IVEA's visual interface

Hence, each document in a text collection can be represented by a record consisting of a number of variables (attributes) whose values correspond to the relevance values of that document with respect to the instances in the personal ontology. The number of variables is called the dimensionality of the data set [10] and an intuitive multi-dimensional visualization component is called for.

3.2 Visual interface and interactions

The visual interface of IVEA is shown in Figure 1. Interested readers are encouraged to view the demo screencast available online[2]. In comparison with the initial version of IVEA, the overview display (the top-right component) is shown on a matrix-based multi-dimensional component instead of a scatter plot. Another improvement on IVEA is the inclusion of a preview panel containing a document viewer which is simple yet sufficient to enable the users to conveniently read the contents of documents from right within IVEA.

The design of the proposed overview display is inspired by the interactive table view used in the FOCUS system, which was originally designed for product comparison and selection [20]. FOCUS can display case-by-attribute tabular data in an intuitive and flexible manner, especially by employing the hierarchical outliner for large attribute sets. While the data set used in FOCUS consists of a mixture of textual and numerical attributes, in our overview display component we only have to deal with numerical attributes (relevance values). As such, our display can be considered as a matrix in which (1) the columns represent documents in a collection, (2) the rows represent entities of interest to the users, and (3) each cell shows how relevant a document is with respect to a class or an instance. We tightly couple the personal interests with this overview display via the hierarchical structure on the left hand side of the matrix as well as on our method to derive the relevance data set as described

[2] http://smile.deri.ie/projects/ivea/demo/demo2.html

in Section 3.1.

We use simple graphics in the cells of the matrix view to communicate relevance values. A fixed color palette serves as a color map to assign a color to each concept or instance and is consistently used throughout all views of IVEA's visual interface. A cell in our component displays a colored bar whose color is determined by the corresponding entity in the ontology it refers to based upon the above-mentioned color map, and whose height is determined by its relevance value. We employ a simple relevance-height mapping scheme in which: (1) the relevance range is from 0 to the highest relevance value in the data set and (2) that range is divided into three intervals, (3) a height value is assigned for each interval. Furthermore, hovering the mouse over a cell displays a tooltip text indicating the associating document name based on its column, the class or instance based on its row, and the relevance value of that document with respect to that entity.

Unlike FOCUS, with our component the users are not provided with the values of all attributes by default. They have total control over which particular entities in their spheres of interest they want to focus on and then visualize the relevance data accordingly. This can be achieved simply by dragging either a class or an instance from the personal knowledge view and dropping it to the overview display area. Since hierarchical relationships are taken into account, dragging and dropping a class over to the overview display will result in the inclusion of all of its direct instances and recursively, all of its subclasses. For example, in Figure 1, the users can drag the class "*pimo:Organization*" from the personal knowledge view and drop it onto the overview display. In this case, all organizations that are defined as instances of "*pimo:Organization*" are automatically included on the hierarchy on the left hand side. The users can gain understanding about which organizations of interest each document mentions by looking at the columns, as well as the distribution of organization entities across documents by looking at the rows. Besides, the users can use the aggregated relevance values for the class "*pimo:Organization*" to see which document is most relevant to all the organizations they are concerned about. The users can of course inspect a text collection along a combination of different dimensions by selecting a set of different classes and instances as shown in Figure 1.

In addition, the users can also dynamically filter out attributes that they are no longer interested in while exploring by just right-clicking on a concept or an instance and the whole row is removed from the overview display. This interaction in effect restricts the matrix to display a particular subset of the data. Furthermore, the hierarchical relationship between elements of the personal ontology allows the users to inspect a text collection at different levels of granularity by expanding or collapsing an expandable node representing a class on the hierarchical structure. These particular forms of semantics-based drilling-down and rolling-up operations allow the users to quickly switch back and forth between abstract and more detailed views.

In the proposed component, we also provide a *crosshair highlighter*. As the users move the mouse over a cell, it highlights both the column and the row containing the cell being hovered. This particular form of spontaneous vertical and horizontal highlighting helps the users to (1) easily recognize which entities of interest are mentioned in a document, together with the details about their relevance values, and (2) the distribution of relevant documents in a collection with respect to a particular entity.

In order to enhance the visual display of the overview component, we employ the barycenter heuristic for edge crossing minimization in bipartite graph drawing research [22] to re-arrange elements of a matrix so that the visual display of data becomes easier to understand. This approach has been previously proposed in [12] to highlight clusters in matrix display. In this approach, the rows and columns are repeatedly reordered in turn based on their barycenters. The barycenter of a column *j* in a matrix takes the value of:

$$\frac{\sum_{i=1}^{n} i a_{ij}}{\sum_{i=1}^{n} a_{ij}}$$

whereby *n* is the number of rows and a_{ij} is the i^{th} entry in column *j* [22]. However, unlike [12], in our work, we only permute the columns instead of permuting both columns and rows, because only the documents are independent of each other and hence the columns can be freely permuted. The entities, on the other hand, are constrained by their hierarchical relationships and thus the rows need to be kept in a certain order. This reordering effort helps to re-arrange the positions of columns of the matrix such that cells with positive relevance values appear more densed in the top left and bottom right corners of the matrix. As a result, it helps the users to interpret the data easier when they appear in groups instead of being scattered over the matrix. Interested readers are referred to [22, 12] for further details about the algorithm.

Lastly, the overview display component is coordinated with the other views in such a way that hovering the mouse over a column in the matrix makes the bar charts, the preview panel, and the TileBars display the corresponding details of the document represented by that column, and hence the users can further investigate other aspects of that document.

4 Discussion

It is inevitable that our proposed approach of tightly coupling a personal ontology defining the users' sphere of interest into a multi-dimensional visualization component for

exploring text collections requires a certain level of efforts from the users to manually build a personal ontology from scratch. Recent study has highlighted five important aspects of building ontologies [7]. Of importance to IVEA among those are the *resource consumption* and the *conceptual dynamics* aspects.

In terms of *resource consumption*, the users gauge whether the benefits gained can outweigh the resources consumed to create an ontology [7]. It is our belief that IVEA can demonstrate that it provides immediate benefits by taking advantage of the defined personal ontology to provide the users with a personal lens to look at text collections from different perspectives. Furthermore, we proposed in [23] the use of the PIMO (Personal Information Model) ontology [17], which serves as a formal representation of concepts and structures within a user's mental model, to minimize the cost of creating a personal ontology. The users only need to extend this readily built ontology with concepts and instances of interest to them. Not only can the users use the PIMO ontology for IVEA, they can benefit from other Semantic Desktop applications using the PIMO ontology as a glue connecting different desktop applications.

With respect to *conceptual dynamics*, it is important that the personal ontology can reflect the quickly evolving domain [7], or in this case, the quickly evolving spheres of interest. With IVEA, the users are provided with the capability to enrich their personal ontologies on-the-fly with new entities matching their interests. This can easily be achieved with minimal effort from the users simply by dragging-and-dropping one of the suggested top frequent phrases from the frequent phrase bar chart onto a tree node in the personal knowledge view. As mentioned earlier, with the newly added entities, the PIMO ontology becomes a richer and better representation of the users' interests and hence can lead to more personalized subsequent exploration and analysis experiences.

In summary, IVEA provides not only intuitive visual components relevant to the exploration task but also other helpful utilities to keep the required efforts from the end-users to a minimum level.

5 Conclusions and Future Work

In this paper, we have reported on an interactive multi-dimensional visualization component which is tightly coupled with the users' personal ontology to allow the users to explore a text collection along different conceptual dimensions. This component is a matrix-based display inspired by the design of the interactive table used in the FOCUS system [20]. We further improve the visual display of this component by applying the barycenter heuristic for edge crossing minimization in bipartite graph drawing to reorder the columns in the matrix. This component is part of IVEA, an information visualization tool which leverages upon a personal ontology representing the users' spheres of interest and employs coordinated and multiple views to highlight various perspectives of documents within a text collection. Of importance in IVEA are (1) the utilization of personal knowledge structure to allow the users to interactively explore a text collection at different levels of granularity and (2) a mechanism to allow the users to incrementally enrich their spheres of interest in the process with minimal effort.

In future work, we plan to enhance the arrangement of the multi-dimensional view, i.e. to re-arrange both sides of the matrix but still maintain the hierarchical structure of the personal ontology. We also intend to improve the personal knowledge view so that the users can manually maintain and extend their personal ontologies. Once all the improvements are made, we will proceed to design and carry out an insight-based evaluation [14] to gauge how IVEA can help the users in gaining insights from exploring a text collection.

Acknowledgments

This work is supported by the Science Foundation Ireland (SFI) under the DERI-Lion project (SFI/02/CE1/1131) and partially by the European Commission 6th Framework Programme in context of the EU IST NEPOMUK IP - The Social Semantic Desktop Project, FP6-027705.

References

[1] M. Ankerst, D. A. Keim, and H.-P. Kriegel. Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets. In *Proceedings of Visualization 96, Hot Topic Session*, San Francisco, CA, 1996.

[2] A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant. Discovering interesting usage patterns in text collections: integrating text mining with visualization. In *CIKM '07: Proceedings of the 16th ACM Conference on information and Knowledge Management*, pages 213–222, New York, NY, USA, 2007. ACM.

[3] C. Fluit. AutoFocus: Semantic Search for the Desktop. In *IV'05: Proceedings of the Ninth International Conference on Information Visualisation*, pages 480–487, Washington, DC, USA, 2005. IEEE Computer Society.

[4] B. Fortuna, M. Grobelnik, and D. Mladenic. Visualization of Text Document Corpus. *Informatica*, 29(4):497–504, 2005.

[5] E. Hatcher and O. Gospodnetic. *Lucene in Action*. Manning Publications Co., Greenwich, CT, USA, 2004.

[6] M. A. Hearst. TileBars: visualization of term distribution information in full text information access. In *CHI '95: Proceedings of the SIGCHI Conference on Human factors in computing systems*, pages 59–66, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co.

[7] M. Hepp. Possible Ontologies: How Reality Constrains the Development of Relevant Ontologies. *IEEE Internet Computing*, 11(1):90–96, 2007.

[8] A. Inselberg and B. Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *VIS '90: Proceedings of the 1st Conference on Visualization '90*, pages 361–378, Los Alamitos, CA, USA, 1990. IEEE Computer Society Press.

[9] E. Kandogan. Star Coordinates: A Multi-dimensional Visualization Technique with Uniform Treatment of Dimensions. In *Proceedings of IEEE Information Visualization, Hot Topics*, pages 4–8, 2000.

[10] D. A. Keim. Information Visualization and Visual Data Mining. *IEEE Transactions on Visualization and Computer Graphics*, 08(1):1–8, 2002.

[11] J. LeBlanc, M. O. Ward, and N. Wittels. Exploring N-dimensional databases. In *VIS '90: Proceedings of the 1st conference on Visualization '90*, pages 230–237, Los Alamitos, CA, USA, 1990. IEEE Computer Society Press.

[12] E. Mäkinen and H. Siirtola. The Barycenter Heuristic and the Reorderable Matrix. *Informatica*, 29(3):357–364, 2005.

[13] J. Mothe, C. Chrisment, B. Dousset, and J. Alaux. DocCube: multi-dimensional visualisation and exploration of large document sets. *Journal of the American Society for Information Science and Technology*, 54(7):650–659, 2003.

[14] C. North. Toward Measuring Visualization Insight. *IEEE Computer Graphics and Applications*, 26(3):6–9, 2006.

[15] R. Rao and S. K. Card. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *CHI '94: Proceedings of the SIGCHI Conference on Human factors in computing systems*, pages 318–322, New York, NY, USA, 1994. ACM.

[16] H. Reiterer, G. Mussler, T. Mann, and S. Handschuh. INSYDER - An Information Assistant for Business Intelligence. In *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR 2000 Conference on Research and Development in Information Retrieval*, pages 112–119. ACM press, 2000.

[17] L. Sauermann, L. van Elst, and A. Dengel. PIMO - a Framework for Representing Personal Information Models. In T. Pellegrini and S. Schaffert, editors, *Proceedings of I-Semantics' 07*, pages 270–277. JUCS, 2007.

[18] C. D. Shaw, J. M. Kukla, I. Soboroff, D. S. Ebert, C. K. Nicholas, A. Zwa, E. L. Miller, and D. A. Roberts. Interactive Volumetric Information Visualization for Document Corpus Management. *Internationl Journal on Digital Libraries*, 2(2-3):144–156, 1999.

[19] B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *IEEE Visual Languages*, pages 336–343, 1996.

[20] M. Spenke, C. Beilken, and T. Berlage. FOCUS: the interactive table for product comparison and selection. In *UIST '96: Proceedings of the 9th annual ACM symposium on User interface software and technology*, pages 41–50, New York, NY, USA, 1996. ACM.

[21] C. Stolte, D. Tang, and P. Hanrahan. Polaris: A System for Query, Analysis, and Visualization of Multidimensional Relational Databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):52–65, 2002.

[22] K. Sugiyama, S. Tagawa, and M. Toda. Methods for Visual Understanding of Hierarchical System Structures. *IEEE Transactions on Systems, Man, and Cybernetics*, 11(2):109–125, 1981.

[23] V. Thai, S. Handschuh, and S. Decker. IVEA: An Information Visualization Tool for Personalized Exploratory Document Collection Analysis. In *ESWC'08: Proceedings of the 5th European Semantic Web Conference*, Tenerife, Spain, 2008.

[24] W. Zhu and C. Chen. Storylines: Visual exploration and analysis in latent semantic spaces. *Computers & Graphics*, 31(3):338–349, 2007.