Visualizing the research profile of an IT Department

Simos Lazaridis
ATEI of Thessaloniki
Information Technology Department
P.O. Box 141
57400, Thessaloniki, Greece
simos@it.teithe.gr

Manolis G. Vozalis
ATEI of Thessaloniki
Information Technology Department
P.O. Box 141
57400, Thessaloniki, Greece
mans@it.teithe.gr

ABSTRACT

The ability of computers to store data is unlimited. As a result, we have at our service as much data as we want, but at the same time emerges the problem of converting them in useful information and knowledge. There are many cases where this information is required to be perceived directly and quickly. The visualization of data is the way to take advantage of the strongest human sense, in order to achieve our goal. Data visualization is the graphical representation of data, aiming to reveal the complex information at a glance.

In this paper, the visualization of the research field of four professors of ATEI of Thessaloniki is attempted. The general idea of visualization, as well as the implementations of some real examples, is also presented. For the data mining, the tf-idf weighting scheme is used upon the words of a corpus, for the extraction of the important concepts. The creation of a thesaurus including the interesting concepts, is compulsory. The counting of their occurrences is done by writing a program in Java. Finally, the vector array which is created, together with the thesaurus, forms the input for the VOS viewer, which is a program for creating knowledge domain visualization maps.

Keywords

Visualization, VOS viewer, research profile, tf-idf

1. INTRODUCTION

Visualization refers to the representation of data with the help of graphics. Such representations aim in discovering implicit information, complex associations or patterns, not easily discernible in any other way.

There have been a number of studies which have described the use of visualization as a means of enhancing human understanding in computer science. Nan Cao et al. [2] present FacetAtlas, a technique that is able to visualize the relations of complex text collections. It allows users to examine a text corpus from various perspectives, providing tools such as filtering or highlighting. Valdis Krebs [4] shows how data mining and data visualization can be combined, leading to sophisticated pattern recognition. This combination allows the exploration of datasets in a way that can reveal interesting insights into the human behaviours behind them. Jason Chuang et al. [3] present the Stanford Dissertation Browser, a visual tool for exploring Ph.D. theses by topical similarity. The proposed model exploits similarity in text collections and discovers word usage patterns in the data.

In this paper we will use visualization methods in order to capture graphically the research profile of the Information Technology (IT) Department of the Technological Education Institute of Thessaloniki (ATEITh), as expressed through the work published (in Conferences and Journals) by four of its professors. These publications will be analyzed so that representative key words can be chosen and a thesaurus will be formed. The words selected for the thesaurus will later be used to give meaning to the visual result. The data, after being processed accordingly, will be utilized as input in the appropriate tools that will aid the creation of the necessary visualizations. In the end, the generated images will be evaluated.

To achieve this kind of visualization, starting from an idea discussed in Van Eck et al [8], we make use of Knowledge Domain Visualization, a term referring to the construction of concept maps, which are utilized to visualize the structure or the evolution of a scientific field. Specifically, our main purpose was to create a map depicting indicative terms from Computer Science, all of them extracted from the published body of work of 4 professors of the IT Department, and the relations among them. Through this visualization, the research interests of each specific professor could be distinguished, and the way these interests are related with each other would be made clear.

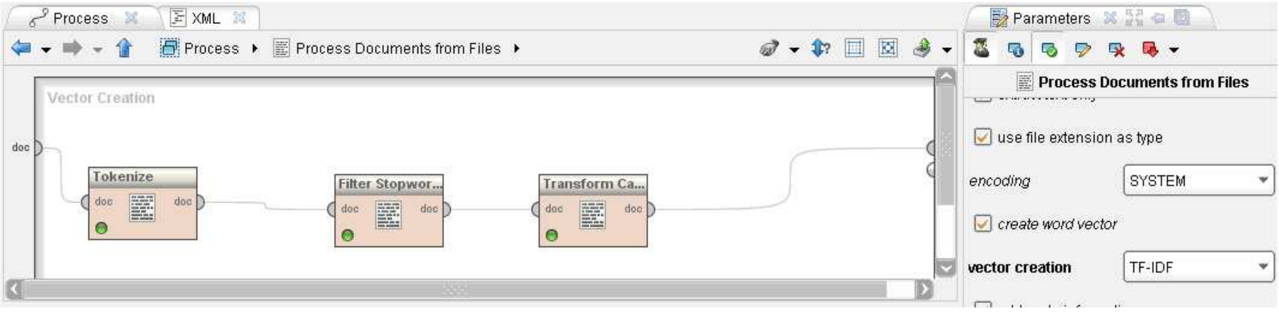
The structure of the paper is this: In Section 2 we give an analytical description of the steps that we followed in order to implement our proposed method. Section 3 provides a discussion of the experiments that we executed, varying specific parameters of the thesaurus, and presents the visualization results. The paper is concluded in Section 4 where also a number of ideas for future research are mentioned.

2. THE IMPLEMENTATION METHOD

In the following paragraphs we will provide a detailed discussion of the steps that were implemented in order to move from the initial raw data to the resulting visualizations.

2.1 Collection of the raw data

What was needed in the first place was a collection of raw



☑ use file extension as type

encoding: SYSTEM

☑ create word vector

vector creation: TF-IDF

Figure 1: Use of RapidMiner for the creation of the thesaurus

data. After the selection of the 4 professors from the IT Department that would be included in our experiments (we will be referring to them with the first letter(s) of their last names, namely "V", "Di", "De" and "S"), we visited their academic web pages and recorded their publications. Based on the fact that only the abstracts of their papers were available in every case, we gathered and stored these abstracts, separately for each professor. The total count of abstracts for all 4 professors was 106.

2.2 Selection of the concept(s) that will be analyzed

For the second step of our implementation we wanted to identify the objects that would be the focal point in our analysis. As mentioned earlier, our intention was to investigate the fields that each of the 4 professors has chosen for his research, and even more importantly, to find how these possibly distinct fields are related to one another. To achieve that, we decided to take into account every possible term seemed to be related to Computer Science, as long as this term appeared in any of the 4 professors' publications - all of them already stored for the previous step of the implementation procedure.

2.3 Creation of the thesaurus

The next step of the implementation procedure involved the creation of a thesaurus including Computer Science related terms. Our plan was to extract nouns or noun phrases from the stored paper abstracts, get rid of stop words - words that are too general or too common to be useful in any way - and sort the remaining terms, based on their significance, by utilizing the *tf-idf weights* that were assigned to them, in a preceding, pre-processing step.

tf-idf[5] takes into account the number of occurrences of a term t in a document d (*term frequency*, $tf_{t,d}$), the total number of occurrences of a term t in a collection of documents c (*collection frequency*, $cf_{t,c}$), the number of documents d in a collection c that includes term t (*document frequency*, df_t), and the total number of documents d included in collection c (D), to first introduce the *inverse document frequency* of a term t as follows,

$$idf_t = log(\frac{D}{df_t}) \qquad (1)$$

and then combine it with term frequency to define tf-idf weighting by:

$$tf_{t,d} = tf_{t,d} \times idf_t \qquad (2)$$

We can now view each document, or abstract in our case, as a vector with one component corresponding to each term in the thesaurus, together with a weight for each component that is given by equation 2. Based on that equation, one can see that the importance of a word is enhanced by the number of times it appears in a document, but that increase is offset by the frequency of the word in the collection of abstracts. Thus, the weight of a term is maximized when it occurs many times in a small number of abstracts (as its presence makes it easier for these abstracts to be distinguished), and minimized when the term occurs in all the stored abstracts (in which case the term bears no discriminating power at all).

For the extraction and sorting of nouns and noun phrases in order to keep the most significant ones, we decided to utilize *RapidMiner*, an open source system ideal for data mining and knowledge discovery[1]. RapidMiner can be used as a stand-alone application for data analysis or as a utility for data mining that can be attached to other products.

In our case, we used the stored abstracts as input to RapidMiner, and tuned it to perform a tokenization of the texts, transform all upper to lowercase letters, and filter out the stop words, based on an existing word list that is retained by the application. A final filter was applied to every abstract file in order to calculate the tf-idf weight of each word appearing in them (Figure 1). The output of RapidMiner was 4 lists, one list corresponding to each professor, that included the selected terms, sorted in tf-idf weight order.

At this point, we decided that each one of the 4 IT professors would contribute proportionally to the thesaurus, based on the count of his abstracts. That is, if a is the count of the abstracts recorded for a specific professor and th is the size of the thesaurus selected for our experiment, his contribution to the thesaurus would be the t terms with the highest tf-idf weight, extracted from the list of his abstracts, where:

$$t = \frac{th * a}{106} \qquad (3)$$

106 is the total number of abstracts among all 4 professors.

Table 1 displays the percentage of terms in the thesaurus, regardless of its size, corresponding to each professor, as calculated by the count of their abstracts.

Figure 2 depicts the whole procedure from the storing of the professors' abstracts to the generation of the thesaurus.

2.4 Extraction of the co-occurency vectors and matrix

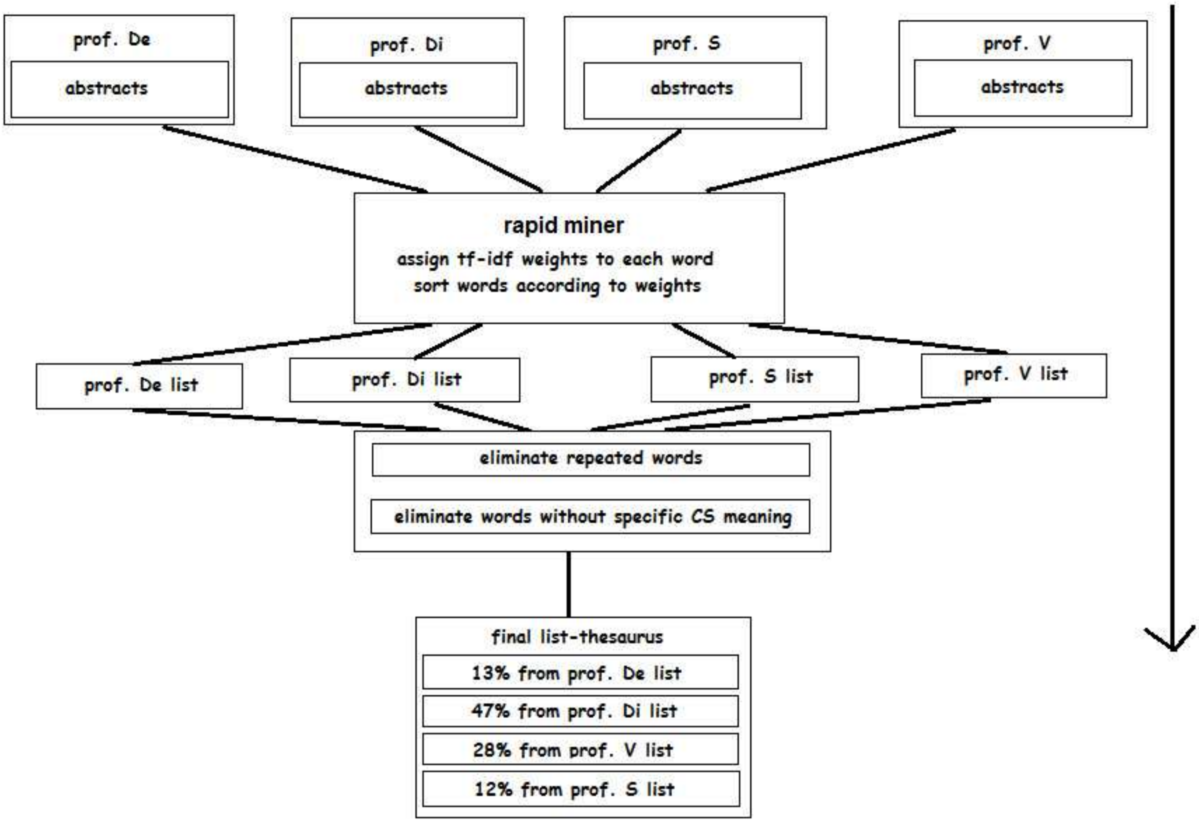


Figure 2: The basic steps of our implementation

Table 1: Percentage contribution of the 4 professors

professor	percentage
V	28%
Di	47%
De	13%
S	12%

Taking into account the thesaurus which was generated in the previous step of the procedure, we should be able to find out when each word from the thesaurus co-exists with every other word in an abstract, and create the corresponding co-occurency vectors.

For that reason, we have written a program in Java that reads all the stored abstracts and looks for words that appear in the thesaurus. Its output consists of two text files: the first file includes the co-occurency matrix, C, with size $n \times n$ (n is the number of terms selected to be included in the thesaurus). Each element of this matrix, $c_{i,j}$, is equal to the number of abstracts that terms i and j, both taken from the thesaurus, appear together. Obviously, C is a symmetric matrix, as $c_{ij} = c_{ji}$. The second file includes all the items (terms) that will be visualized. The rows of the "items file" are equal to the rows (or columns) of the "co-occurency file".

2.5 Visualization

The positioning of the different items in a low-dimensional space, based on their similarities, is achieved by the application of a method called VOS, which is an abbreviation for *visualization of similarities* [6]. In our case, the items represent Computer Science concepts. A computer program called VOSviewer [7] is run to place these concepts in a 2-dimensional space, by utilizing the VOS technique. It will create the corresponding concept map by taking into account the co-occurency and the item files, both generated in the previous step of the procedure.

3. EXPERIMENTS AND RESULTS

In this section we will focus on some important implementation details. We will discuss how the size of the thesaurus affects the output and run different experiments by varying it in order to compare the resulting visualizations.

3.1 The base-case experiment

For our base-case experiment we chose a thesaurus with 100 words. No term selection was implemented and all words included were chosen solely from their calculated weights.

Based on our decision to allow each professor to contribute proportionally to the thesaurus and the percentages shown in Table 1, it can be concluded that professor V would contribute to the 100-word thesaurus with 28 terms, professor Di with 47 terms, professor De with 13 and professor S with 12 terms.

Figure 3 shows the visualization of the thesaurus including

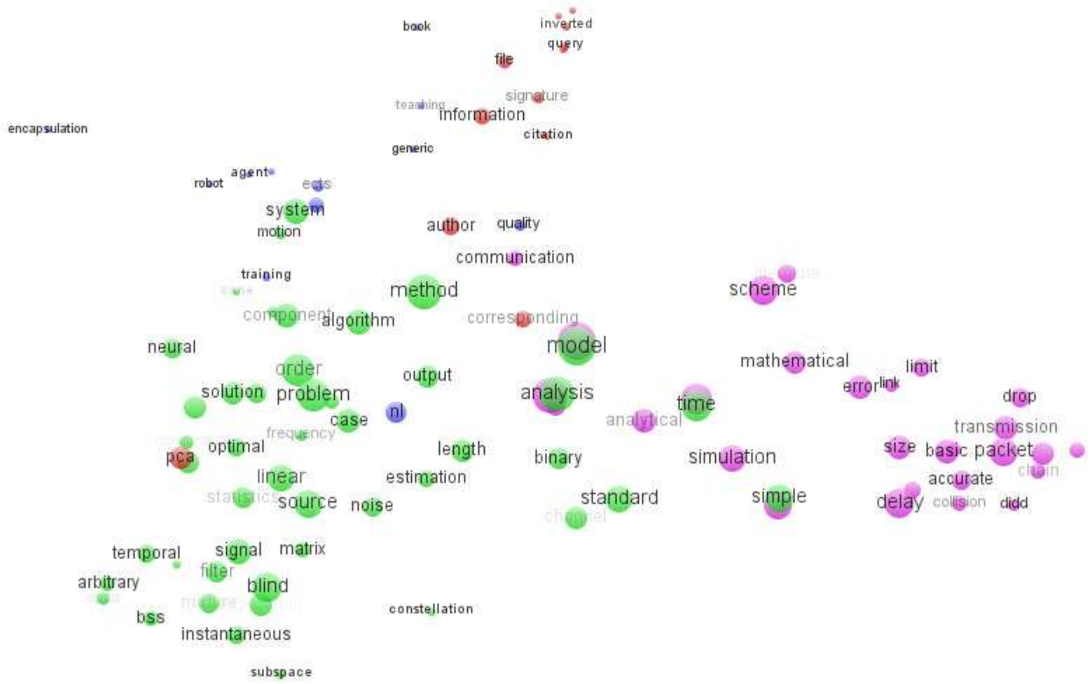


Figure 3: A Visualization of a thesaurus including 100 terms

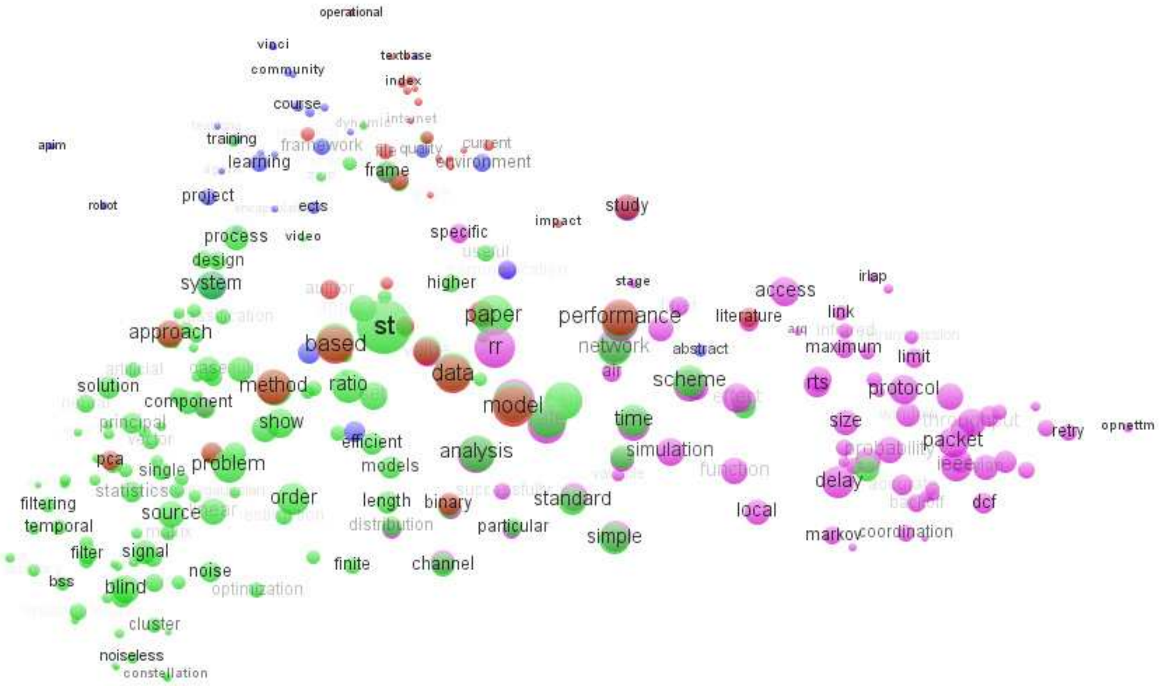


Figure 4: A Visualization of a thesaurus including 300 terms

100 terms. Each colored "bubble" - with every color corresponding to a different professor - depicts a distinct term from the thesaurus, with its size depending on the term's frequency: the more abstracts a term appears in, the bigger the "bubble" it is represented by. As a result, one might claim that the size of a "bubble" can be viewed as a measure for the term's importance.

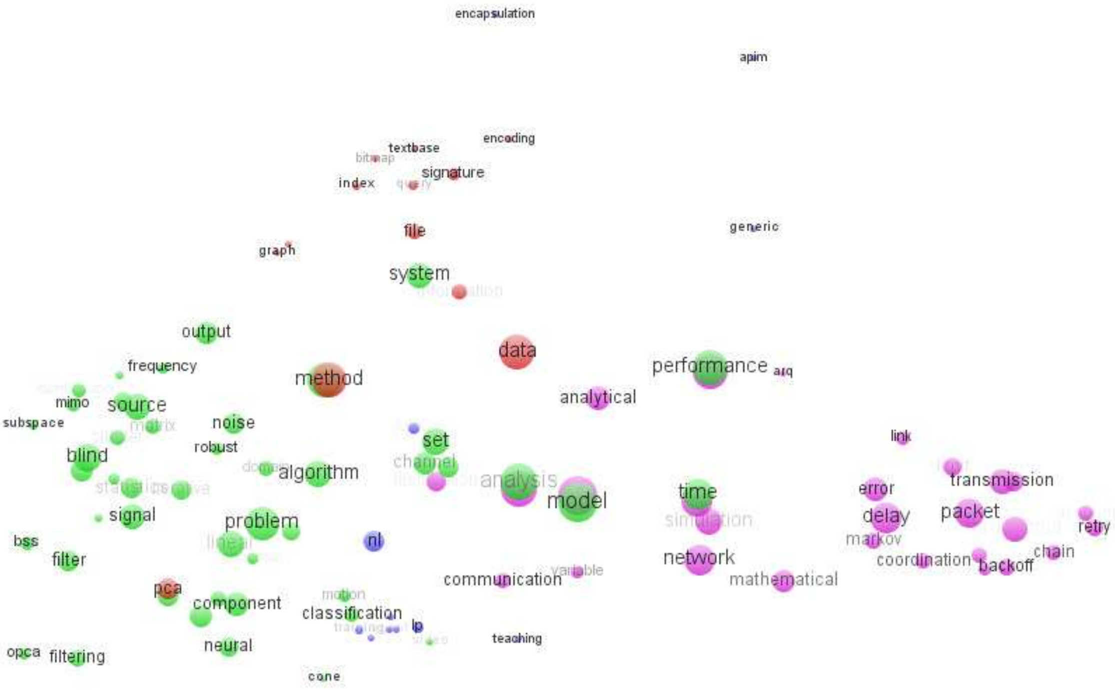


Figure 5: A Visualization of an enhanced thesaurus including 100 terms

3.2 More experiments

Accordingly, we run experiments with varying thesaurus sizes in order to compare the visualization results. Specifically, we run experiments that involved thesauruses including 50, 150, 200, 300 and 1000 terms. Again, in all cases, no special word selection was made. Table 2 records the number of terms that each professor contributed to these thesauruses.

Table 2: Absolute contribution of the 4 professors

Thesaurus Size	50	150	200	300	1000
V	14	42	56	84	280
Di	23	70	94	141	470
De	7	20	26	39	130
S	6	18	24	36	120

Among all experiments that were executed and their corresponding results, due to space limitation, we present here only one. Figure 4 depicts the visualization of the thesaurus with 300 terms.

It is interesting to note that the increase in the size of the thesaurus gives a richer, more colorful image. At the same time, it is important to strike the right balance in that increase: a more dense image may not necessarily aid the reader in making the desired observations, based on the generated visualization of the initial data. It appears that the 300 term image lies on the verge of being considered an "overloaded" one.

3.3 Experiments with an enhanced thesaurus

For our next set of experiments, we took the original thesauruses including 100 and 150 terms, removed words without any apparent importance, and replaced them with others, of lower tf-idf weight, selected because of their relation to Computer Science (CS). We will be referring to the generated thesauruses and corresponding visualizations as "enhanced".

In Figure 5 we are presenting the first of the two experiments of that series, which depicts the visualization of the enhanced thesaurus with 100 terms. The significance of the enhanced visualization lies in the fact that after the CS term selection, even a more limited, in terms of included words, image gives the impression of a denser one, with the research areas of the 4 professors and their relationships more easily discernable: Professor Di's terms (green "bubbles"), expressing his work in Neural Networks' related areas, are placed to the left side of the image, while professor V's "bubbles", colored purple, expressing his work in Networks, to the right side. The terms from professors De (red "bubbles") and S (blue "bubbles") were placed somewhere between the two, but clearly closer to the area attributed to professor Di. This placement confirms the established fact that their research interests (data/text mining, artificial intelligence etc.) lie closer to the interests of professor Di, than to those of professor V.

3.4 Some visual observations

Figures 6 and 7 show the visualizations of the original and the enhanced thesauruses, with both including 100 terms. Their main difference when compared to preceding Figures (3 and 5) is that the clusters with the terms corresponding to each of the 4 professors have been placed in bounding boxes. Taking into account that the research field of a participating professor is expressed visually through the terms taken from his abstracts, the presence of the bounding boxes helps us

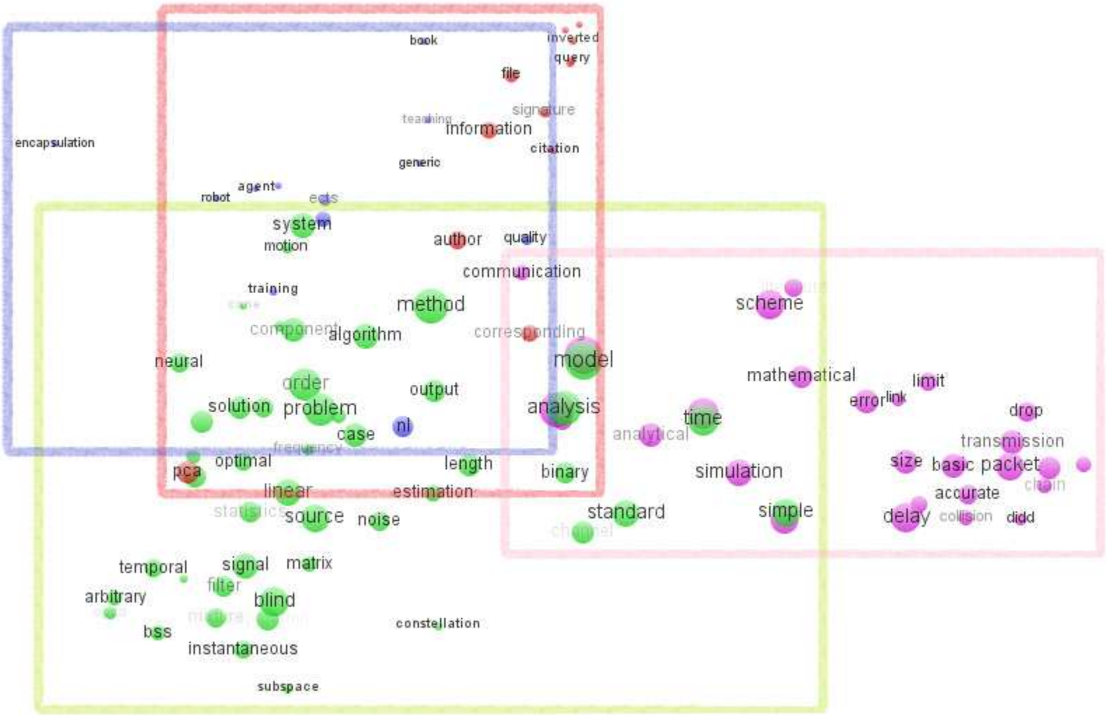


Figure 6: Visualization of the original 100-term thesaurus with bounding boxes

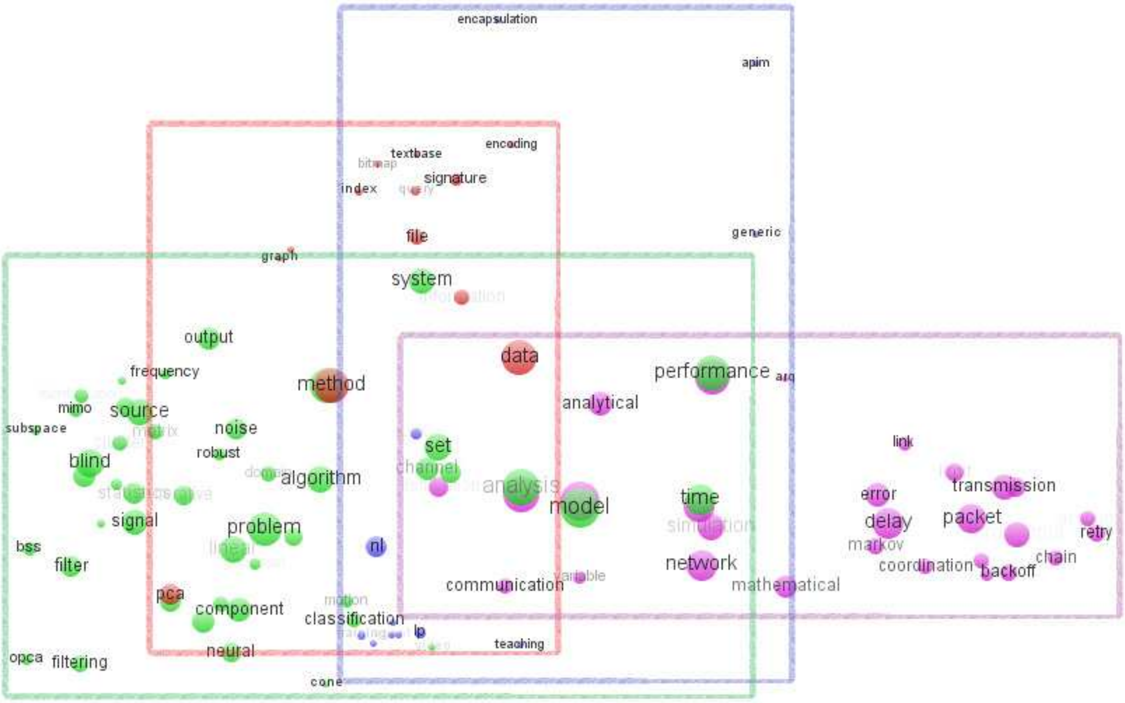


Figure 7: Visualization of the enhanced 100-term thesaurus with bounding boxes

reach conclusions about the correlation of the professors' scientific interests in an easier manner.

Figure 6 gives us an initial idea about the possible relations among the 4 professors. The greater differences are located in the fields of professors Di and V, fittingly lying at the opposite ends of the image. At the same time, professors' De and S boxes lie in the middle and almost coincide. This observation constitutes the biggest problem of the specific visualization, because the suggested relation, as expressed by this placement, does not correspond to the reality.

Once terms with little or no meaning are removed and replaced with CS words (Figure 7) we achieve an image where the placement of the 4 boxes becomes clearer and, more importantly, closer to what is true. Once more, the boxes that belong to professors Di and V occupy the left and the right borders, respectively. This time, though, the boundaries between professors De and S are comprehensible, with both their boxes leaning slightly to the left, expressing a closer relation to professor Di.

Conclusively, the research fields of the 4 professors are more accurately represented in Figure 7, making the specific image an improvement over the previous one. This observation lets us conclude that, apart from the size, the quality of the selected thesaurus plays a decisive role in the effectiveness of the visualization result.

4. CONCLUSIONS AND FUTURE WORK

In this paper we described a method for the visualization of the research profile of the IT Department of the Technological Education Institute of Thessaloniki, as expressed by the fields of research of four of its professors. Our purpose was to take words, preferably computer science terms, from a pre-defined thesaurus that was compiled from the professors' scientific publications, and place them in the 2-dimensional space in a way that would allow us to draw certain conclusions. In the end we were able to reach interesting conclusions about the frequency and the significance of the selected terms and, more importantly, about the relation and the proximity of the research interests of the participating professors.

Future work may involve the utilization of the latest VOS viewer edition which, among others, features the use of text mining techniques. Another meaningful addition would be the introduction of the concept of time, in order to check how it would affect the generated visualizations. A continually changing concept map where some terms would become more eminent while others could even disappear as time passes, would be of great usefulness, giving us a clear idea of the way research in specific scientific fields is evolving. Furthermore, it would be interesting to add more professors from the IT Department in the mix, expanding our concept maps and possibly locating different relations or clusters that involve their research interests. Finally, we could try alternative ways of calculating the professors' contribution in the visualization results, other than determining it strictly proportionally to the count of their abstracts.

5. REFERENCES

[1] Rapidminer. `http://rapid-i.com/content/view/181/190/`. Accessed February 23, 2013.

[2] N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu. Facetatlas: Multifaceted visualization for rich text corpora. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1172–1181, November 2010.

[3] J. Chuang, D. Ramage, C. D. Manning, and J. Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 443–452, Austin, Texas, USA, 2012.

[4] V. Krebs. Your choices reveal who you are: Mining and visualizing social patterns. In J. Steele and N. Iliinsky, editors, *Beautifu Visualization, Looking at Data through the eyes of Experts*, pages 103–122. O'Reilly, 2010.

[5] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, New York, 2008.

[6] N. J. van Eck and L. Waltman. VOS: a new method for visualizing similarities between objects. In *Proceedings of the 30th Annual Conference of the German Classification Society*, pages 299–306, 2007.

[7] N. J. van Eck and L. Waltman. VOSviewer: A computer program for bibliometric mapping. In *Proceedings of the 12th International Conference on Scientometrics and Informetrics*, pages 886–897, 2009.

[8] N. J. van Eck, L. Waltman, J. van den Berg, and U. Kaymak. Visualizing the WCCI 2006 knowledge domain. In *Proceedings of the 2006 IEEE International Conference on Fuzzy Systems*, pages 7862–7869, Vancouver, BC, Canada, 2006.