Visualization of text document corpus

Blaž Fortuna, Marko Grobelnik, Dunja Mladenić

Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
{blaz.fortuna, marko.grobelnik, dunja.mladenic}@ijs.si

Keywords: text visualization, latent semantic indexing, mutlidimensional scaling

Received: November.2005

Visualization is commonly used in data analysis to help the user in getting an initial idea about the raw data as well as visual representation of the regularities obtained in the analysis. In similar way, when we talk about automated text processing and the data consists of text documents, visualization of text document corpus can be very useful. From the automated text processing point of view, natural language is very redundant in the sense that many different words share a common or similar meaning. For computer this can be hard to understand without some background knowledge. We describe an approach to visualization of text document collection based on methods from linear algebra. We apply Latent Semantic Indexing (LSI) as a technique that helps in extracting some of the background knowledge from corpus of text documents. This can be also viewed as extraction of hidden semantic concepts from text documents. In this way visualization can be very helpful in data analysis, for instance, for finding main topics that appear in larger sets of documents. Extraction of main concepts from documents using techniques such as LSI, can make the results of visualizations more useful. For example, given a set of descriptions of European Research projects (6FP) one can find main areas that these projects cover including semantic web, e-learning, security, etc. In this paper we describe a method for visualization of document corpus based on LSI, the system implementing it and give results of using the system on several datasets.

1 Introduction

Automated text processing is commonly used when dealing with text data written in a natural language. However, when processing the data using computers, we should be aware of the fact that many words having different form share a common or similar meaning. For a computer this can be difficult to handle without some additional information -- background knowledge. Latent Semantic Indexing (LSI) is a technique for extracting this background knowledge from text documents. It employs a technique from linear algebra called Singular Value Decomposition (SVD) and the bag-of-words representation of text documents for extracting words with similar meanings. This can also be viewed as the extraction of hidden semantic concepts from text documents.

Visualization of a document corpus is a very useful tool for finding the main topics that the documents from this corpus talk about. Different methods were proposed for visualizing a large document collection using different underlying methods. For instance, visualization of large document collection based on document clustering [3], or visualization of news collection based on visualizing relationships between named entities extracted from the text [4]. Another example used in our work is visualization of European research space [5]. Given a set of descriptions of European research projects in IT (6th Framework IST), using document visualization one can find main areas that these projects cover, such as *semantic web, e-learning, security,* etc.

In automated text processing document are usually represented using the bag-of-words document representation, where each word from the document vocabulary stands for one dimension of the multidimensional space of documents. Consequently, in automated text processing we are dealing with very high dimensionality of up to hundreds of thousands dimensions. Dimensionality reduction [6] is important for different aspects of automated text processing including document visualization.

We propose to use dimensionality reduction for document visualization by first extracting main concepts from documents using LSI and than using this information to position documents on a two dimensional plane via multidimensional scaling [1]. The final output is graphical presentation of a document set that can be plotted on a computer screen. The proposed approach is implemented as a part of *Text Garden* software tools for text mining[7][1] in a component providing different kinds of document corpus visualization based on LSI and multidimensional scaling.

This paper is organized as follows. Section 2 provides a short description of LSI and multidimensional scaling, while its application to document visualization is given in Section 3. Description of the developed system implementing the method is given in Section 4. Section 5 provides conclusions and discussion.

[1] `http://www.textmining.net/`

2 Building blocks

First step of our approach to visualization of a document corpus is mapping all the documents into two dimensional space so we can plot them on a computer screen. Ideally they would be positioned in such a way that the distance between two documents would correspond to the content similarity between them.

We obtain this mapping by sending the document corpora trough the pipeline for reducing dimensionality, consisting from building blocks presented in this Section. The whole pipeline will be outlined in the Section 3.

2.1 Representation of text documents

The first step in our approach is to represent text documents as vectors. We use the standard Bag-of-Words (BOW) representation together with TFIDF weighting [9]. In the BOW representation there is a dimension for each word; a document is encoded as a feature vector with word frequencies as elements. Elements of vectors are weighted, in our case using the standard TFIDF weights as follows. The i-th element of the vector containing frequency of the i-th word is multiplied with $IDF_i = \log(N/df_i)$, where N is total number of documents and df_i is document frequency of the i-th word (the number of documents from the whole corpus in which the i-th word appears).

2.2 Latent Semantic Indexing

A well known and used approach for extracting latent semantics (or topics) from text documents is Latent Semantic Indexing [2]. In this approach we first construct term-document matrix A from a given corpus of text documents. This is a matrix with vectors of documents from a given corpus as columns. The term-document matrix A is then decomposed using singular value decomposition, so that $A = USV^T$; here matrices U and V are orthogonal and S is a diagonal matrix with ordered singular values on the diagonal. Columns of matrix U form an orthogonal basis of a subspace in the bag-of-words space where vectors with higher singular values carry more information -- this follows from the basic theorem about SVD, which tells that by setting all but the largest k singular values to 0 we get the best approximation for matrix A with matrix of rank k). Vectors that form the basis can be also viewed as concepts and the space spanned by these vectors is called the *Semantic Space*.

Each concept is a vector in the bag-of-words space, so the elements of this vector are weights assigned to the words coming from our documents. The words with the highest positive or negative values form a set of words that are found most suitable to describe the corresponding concept.

A related approach (not used here) that also aims at extracting latent semantics from text documents is Probabilistic Latent Semantic Analysis (PLSA) introduced in [8]. Compared to standard Latent Semantic Analysis which comes from linear algebra and performs a Singular Value Decomposition of co-occurrence tables, this method is based on a mixture decomposition derived from a latent class model. This method assigns each word a probability to be in a concept, where the number of concepts is predefined.

2.3 Dimensionality reduction

We are using a sequential combination of linear subspace methods and multidimensional scaling for reducing document space dimensionality. Both methods can be independently applied to any data set that is represented as a set of vectors in some higher dimensional space. Our goal is to lower the number of dimensions to two so that the whole corpus of documents can be shown on a computer screen.

Linear subspace methods [10], like Principal Component Analysis (PCA) or Latent Semantic Indexing, focus on finding direction in original vector space, so they capture the most variance (as is the case for PCA) or are the best approximation for original document-term matrix (as is the case for LSI). By projecting data (text documents) only on the first two directions we can get the points that live in the two dimensional space. The problem with linear subspace methods is that only the information from the first two directions is preserved. In case of LSI it would mean that all documents are described using only the two main concepts.

Multidimensional scaling [1] enables dimensionality reduction by mapping original multidimensional vectors onto two dimensions. Here the points representing documents are positioned into two dimensions so they minimize some energy function. The basic and most common form of this function is

$$E = \sum_{i \neq j} \delta_{ij} - d(x_i, x_j))^2,$$

where x_i are two dimensional points and δ_{ij} represents the similarity between two vectors (in our case documents i and j). An intuitive description of this optimization problem is: the better the distances between points on the plane approximate real similarity between documents, the lower the value of the energy function. Function E is nonnegative and equals zero only when distances between points match exactly with similarity between documents.

3 Visualization using dimensionality reduction

We propose combining the two methods (linear subspace and multidimensional scaling) in order to take advantage of the nice properties they both have. What follows is description of the proposed algorithm:

Input: Corpus of documents to visualize in form of TFIDF vectors.
Output: Set of two dimensional points representing documents.
Procedure:

1. Calculate k dimensional semantic space generated by input corpus of documents,
2. Project documents into the semantic space,
3. Apply multidimensional scaling using energy function on documents with Euclidian distance in semantic space as similarity measure.

There are two main problems to be solved to make the above algorithm work efficiently. First problem is how to determine the value of k. The way we choose is by checking the singular values. Let $\Sigma_k = S_1^2 + S_2^2 + \ldots + S_k^2$, where S_i is i-th singular value. We know that $\Sigma_n = \mathrm{Trace}(A^TA)$, where n is the number of the documents in the corpus and A is the term-document matrix. From this we can guess k by prescribing the ratio Σ_k / Σ_n to some fixed value, eg., 50%.

A more difficult problem is how to perform multidimensional scaling efficiently. Common way is to use gradient descent. The problem with this approach is that the energy function is not convex: it usually has many local minima which are not that interesting for us. One could start this method more times with different initial state and than choose the results with the lowest energy.

We choose a slightly different way which is based on reformulation of the energy function. Given a placement of points, we calculate for each point how to move it so we minimize energy function. Lets denote the current positions of points with (x_i,y_i) and the desired position with $(x_i',y_i') = (x_i + \delta x_i, y_i + \delta y_i)$. Than we have

$$d_{ij}'^2 - d_{ij}^2 = (x_i - x_j)^2 + (y_i - y_j)^2 -$$
$$(x_i + \delta x_i - x_j - \delta x_j)^2 + (y_i + \delta y_i - y_j - \delta y_j)^2 \approx$$
$$\approx (x_i - x_j)\,\delta x_i + (x_j - x_i)\,\delta x_j + (y_i - y_j)\,\delta y_i + (y_j - y_i)\,\delta y_j =$$
$$= [(x_i - x_j), (x_j - x_i), (y_i - y_j), (y_j - y_i)][\,\delta x_i, \delta x_j, \delta y_i, \delta y_j]^T.$$

By writing this for each pair (i,j) and substituting d_{ij}' with the original distance between i-th and j-th document we get a system of linear equations which has a vector of moves (δx and δy) for a solutions. This is an iteration which finds a step towards minimizing energy function and is more

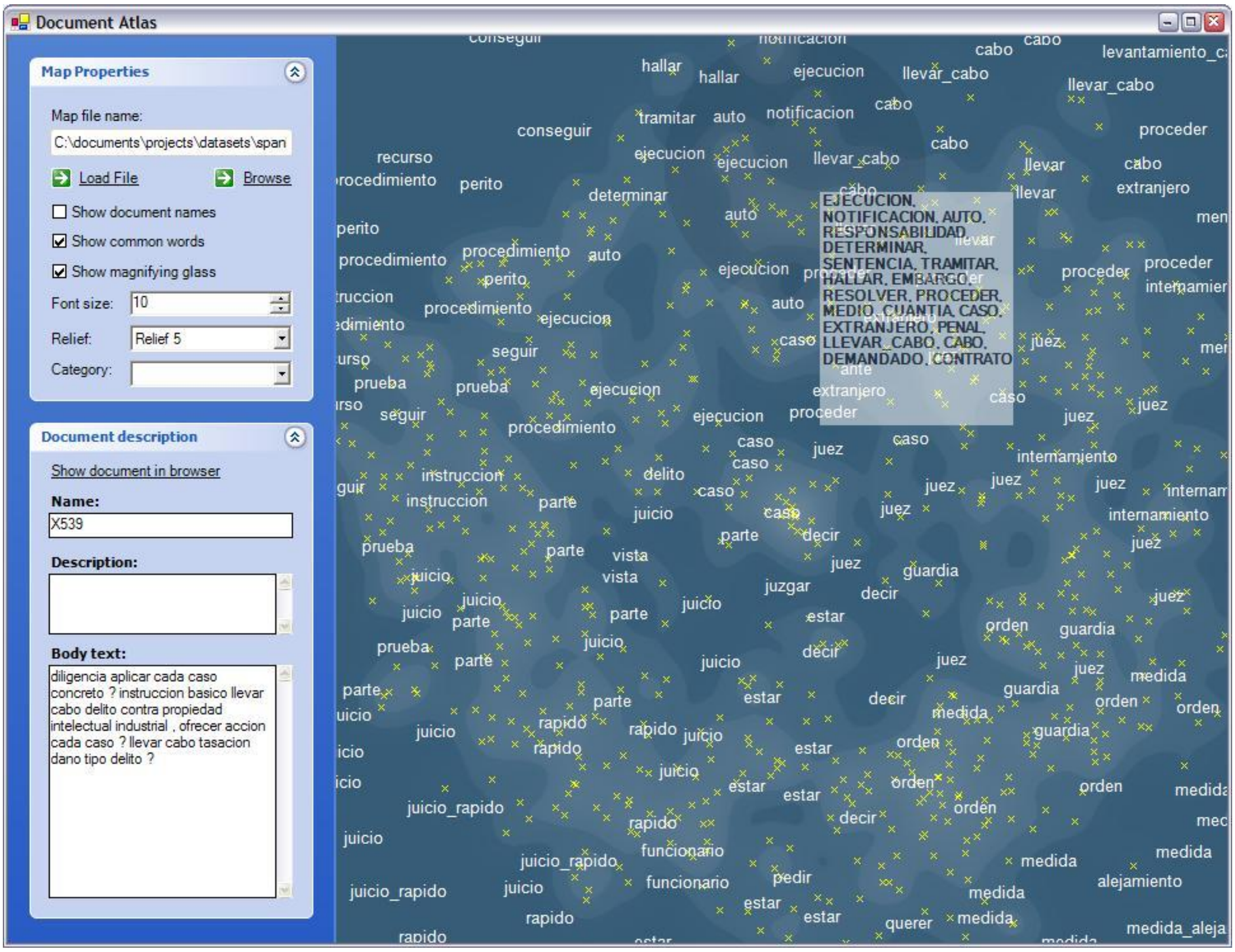

Figure 1 Visualization of questions. The dataset is a collection of around 800 most frequent questions asked by Spanish judges regarding the law and trials. Each question is treated as one document. Dataset is taken from a case study of the EU SEKT project.

successful at avoiding local minima. Each iteration involves solving a linear system of equations with a very sparse matrix. This can be done very efficiently using Conjugate Gradient (CG) method. Finally, the points are normalized to lie in the square $K = [0,1]^2$.

4 Visualization beyond dimensionality reduction

After the documents from corpus are mapped onto a two dimensional plane, some other techniques can be used to make the structure of documents more explicit for the users:

- ***Landscape generation***: landscape can be generated by using the density of points. Each point from square K is assigned height using the formula $h(x, y) = \sum_i \exp(-\sigma \, ||(x,y) - (x_i, y_i)||^2)$.
- ***Keywords***: each point from square K can be assigned a set of keywords by averaging TFIDF vectors of documents which appear within a circle with centre in this point and radius R.

We use these features when showing visualizations to make them more descriptive and to improve the overall user experience.

In our system, called *Text Garden Document Atlas*, the documents are presented as yellow crosses on a map and the density is shown as a texture in the background of the map (the lighter the color, the higher the density). The most common keywords are shown for the areas around the map. Positions, for which keywords are computed, are selected randomly. Keywords are displayed using white color font.

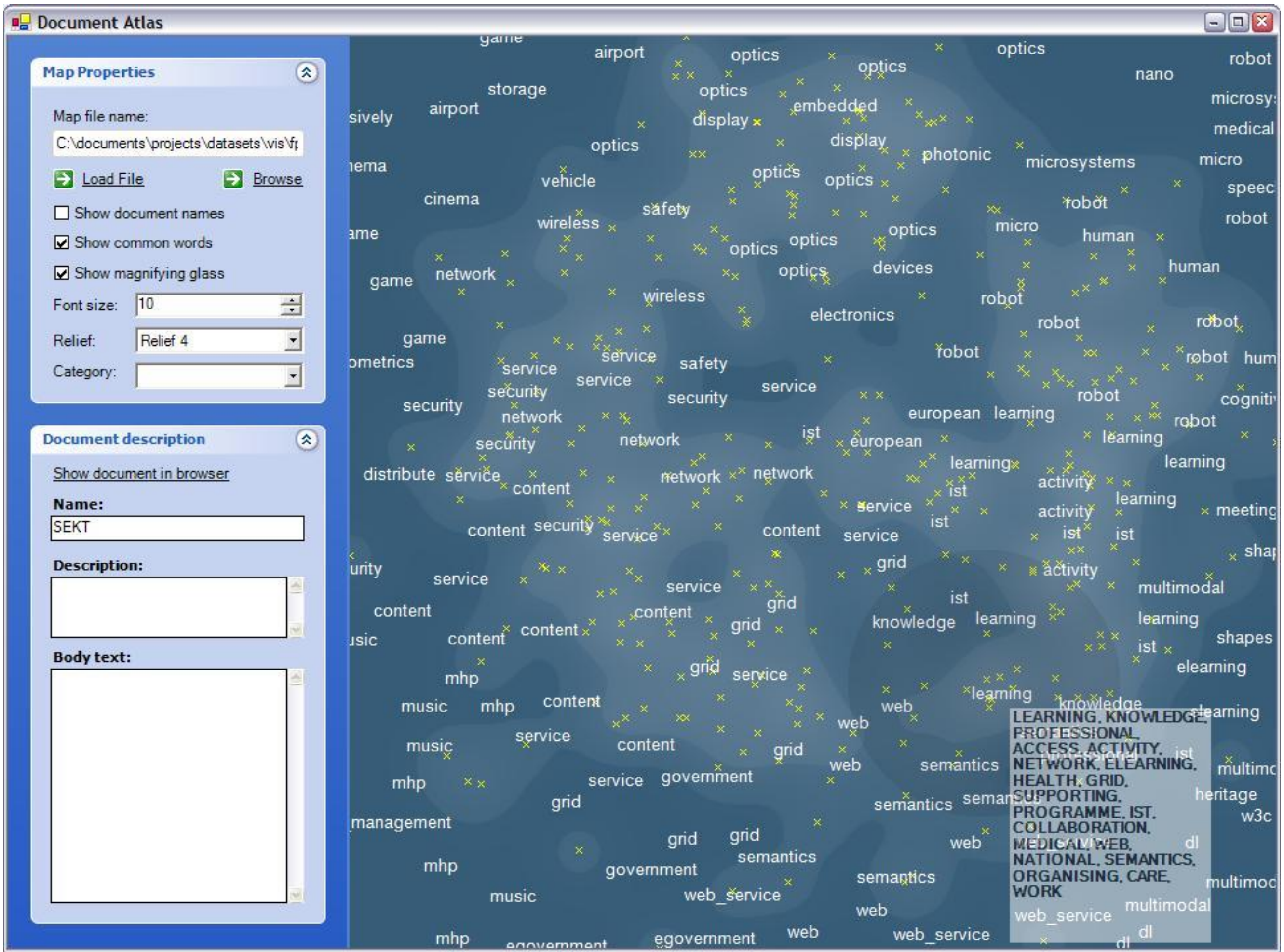

Figure 2 Visualization of European IST projects from 6th framework. Dataset consists of descriptions of EU IST projects taken from CORDIS web site. One can see in the visualization the main areas covered by the projects. The lower right side consists of projects about semantic web. In counter-clockwise direction the topics change to multimodal integration, e-learning, robotics, optics, safety, networking, grid computing, and than back to web related projects. These topics can be easily read from the map by checking the keywords. We can notice that besids putting similar documents together, the visualizations also puts similar topics more close on the map. Each document from the dataset corresponds to a description of one research project.

When the user moves the mouse around the map a set of the most common keywords is computed in real-time for the area around the mouse (the area, from which these keywords are extracted, is marked darker on the map and the list of keywords is shown in the semi-transparent window next to the mouse). The user can also zoom-in to see specific areas in more details. By clicking on a document (yellow crosses on the map), more information about it is shown on the left bottom side side of the screen.

Two examples of the visualizations can be seen in Figure 1 and 2. Figure 1 shows visualization of a corpus with questions from Spanish judges [8]. A map of the European IST projects from 6th framework is shown in Figure 2 and zoom-in on the part showing projects related to web and semantic web is shown in Figure 3..

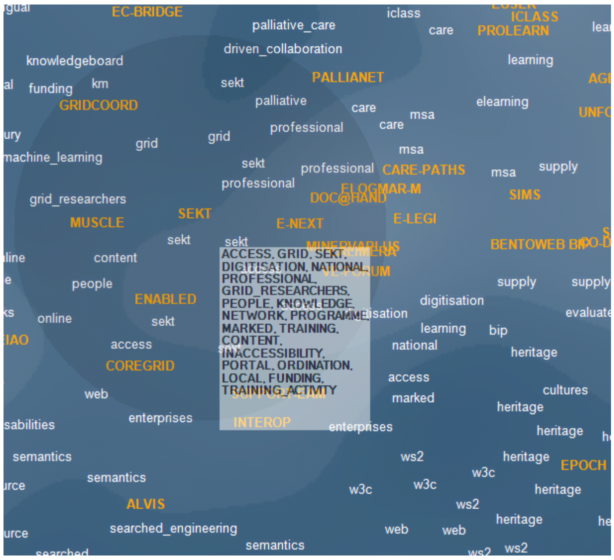

Figure 3 Zoom-in to projects related to web such as 6FP projects SEKT, MUSCLE, ALVIS, etc. Uppercase orange words are the project names (acronyms).

5 Conclusions and future work

We have proposed a approach to efficient visualization of large data collections and describe the developed system implementing the proposed approach. The system was successfully used for visualizing different kinds of document corpora – from project descriptions, scientific articles to short questions from legal domain and even clients of an Internet grocery store. We found that the system is very helpful for data analysis offering quick insight into the structure of the visualized corpus. However providing a more systematic evaluation of the system including users questionnaire remains for the future work.

We will continue to use the user feedback as a guide for adding new features, which would make this tool even more informative and useful. One area not fully explored yet is the use of background relief in visual representation of the document corpus. Currently we use the relief to show the density of documents but it can also be used for showing some other attributes. Another direction we are considering for future work is to improve scalability by making the multi-dimensional step more scalable with the number of documents.

Acknowledgement

This work was supported by the Slovenian Research Agency and the IST Programme of the European Community under SEKT Semantically Enabled Knowledge Technologies (IST-1-506826-IP) and PASCAL Network of Excellence (IST-2002-506778). This publication only reflects the authors' views.

References

[1] Carroll, J.D., Arabie, P. Multidimensional scaling. *In M.R. Rosenzweig and L.W. Porter (Eds.), Annual Review of Psychology*, 1980, 31, 607-649.

[2] Deerwester, S., Dumais, S., Furnas, G., Landuer, T., Harshman, R. Indexing by Latent Semantic Analysis, *Journal of the American Society of Information Science*, 1990.

[3] Grobelnik, M., and Mladenic, D.: *Efficient visualization of large text corpora*. Proceedings of the Seventh TELRI seminar. Dubrovnik, Croatia, 2002.

[4] Grobelnik, M., Mladenic, D. Visualization of news articles. Informatica journal, 2004, vol. 28, no. 4.

[5] Grobelnik, M., Mladenic, D. Analysis of a database of research projects using text mining and link analysis. In *Data mining and decision support : integration and collaboration*, Kluwer Academic Publishers, Boston; Dordrecht; London, 2003, pp. 157-166.

[6] Grobelnik, M., Mladenic, D. Text Mining Recipes, Springer-Verlag, Berlin; Heidelberg; New York (to appear), 2006, (accompanying software available at <http://www.textmining.net>.

[7] Mladenic, D. Feature Selection for Dimensionality Reduction. In *Subspace, Latent Structure and Feature Selection techniques: Statistical and Optimisation perspectives*, Springer Lecture Notes in Computer Science (to appear), 2006.

[8] Hoffman, T., *Probabilistic Latent Semantic Analysis*, Proc. of Uncertainty in Artificial Intelligence, UAI'99, 1999

[9] Salton, G. Developments in Automatic Text Retrieval, *Science*, Vol 253, pages 974-979, 1991.

[10] Shawe-Taylor, J., Cristianini, N. *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004, 143-150

[11] Vallbe, J.J., Marti, M.A., Fortuna, B., Jakulin, A., Mladenic, D., Casanovas, P. Stemming and lemmatisation, Springer Lecture Notes, (to appear), 2006.