VISUALIZATION OF TEXT DOCUMENT CORPUS

Blaž Fortuna, Dunja Mladenić, Marko Grobelnik
Department of Knowledge Technologies
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773419; fax: +386 1 4251038
e-mail: blaz.fortuna@ijs.si

ABSTRACT

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. Latent Semantic Indexing (LSI) is a technique that helps in extracting some of this background knowledge from corpus of text documents. This can be also viewed as extraction of hidden semantic concepts from text documents. On the other hand 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) [2] 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. For example, 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. Bag-of-words representation of text has very high dimensionality, so in order to visually represent text documents, the number of dimensions has to be reduced. This can be done by first extracting main concepts from documents using LSI and than using this information to position documents on a two dimensional plane that can be plotted on computer screen.

As a part of *Text Garden* software tools for text mining[1] we have developed a component that provides different kinds of document corpus visualization based on LSI and multidimensional scaling [3]. 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.

2 BUILDING BLOCKS

2.1 Representation of text documents

In order to use the algorithms, which we will presented below, text documents must first be represented as vectors. We use the standard Bag-of-Words (BOW) representation together with TFIDF weighting [1]. 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

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

Semantic Indexing. 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 famous 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 of rank k). Because of all this, vectors that form the basis can be also viewed as concepts. 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.

2.3 Dimensionality reduction

We are using linear subspace methods and multidimensional scaling as methods for reducing space dimensionality. They can be both applied to any data set that is represented as a set of vectors in some higher dimensional space. Our goal here was 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, 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 this approach 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.

Another approach is called multidimensional scaling [3]. 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 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) as they have some nice properties that fit together. 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. One way of doing this 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 $\Sigma_n = \text{Trace}(A^T A)$, where n is the number of the documents in the corpus and A is the term-document matrix. From this we can guess the k by prescribing the ratio Σ_k / Σ_n to some fixed value, for example 50%.

A more difficult problem is how to perform multidimensional scaling efficiently. One 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. This energy function can also be reformulated in the following way. Given a placement of points, we calculate how to move each point 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

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

By writing this down as a matrix we get a system of linear equations which has a vector of moves towards the minima (δx and δy) for the solution. This is an iteration which finds a step towards minimizing energy function and is more successful at avoiding local minima. Each iteration involves solving a linear system of equations with a sparse

matrix (4 non-zero elements per row). We do this very efficiently using the Conjugate Gradient method. At the end, points are normalized to lie in the square $K = [0,1]^2$.

4 VISUALIZATION BEYOND DIMENSION 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 points on a map and the density is shown as a texture in the background of the map. The most common keywords are shown for the areas around the map. When the user moves the mouse around the map a set of the most common keywords is shown for the area around the mouse (the area, from which these keywords are extracted, is marked with a circle). The user can also zoom-in to see specific areas in more details. By clicking on a document, more information about it is shown on the left side of the screen.

Two examples of the visualizations can be seen in Figure 1 and Figure 2. Figure 1 shows visauliation of the European IST 6th framework projects collected as follow up of 5th framework analysis [4]; here each document from the input corpus is a description of one research project. Figure 2 shows visualization of a corpus with questions from Spanish judges collected through interviews with judges [5]; here one question is treated as one document.

5 CONCLUSION AND FUTURE WORK

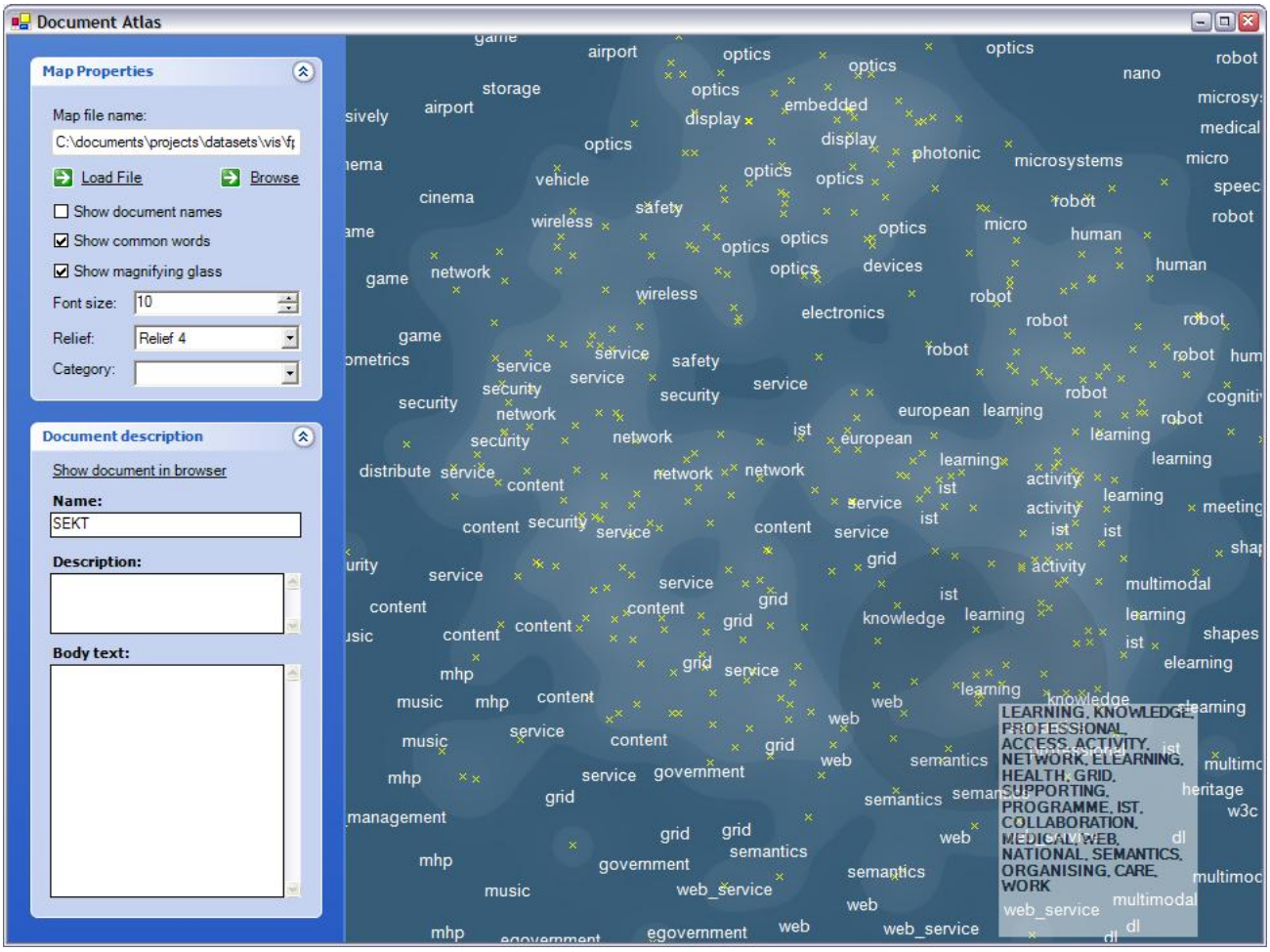

Figure 1. Visualization of European IST projects from 6th framework.

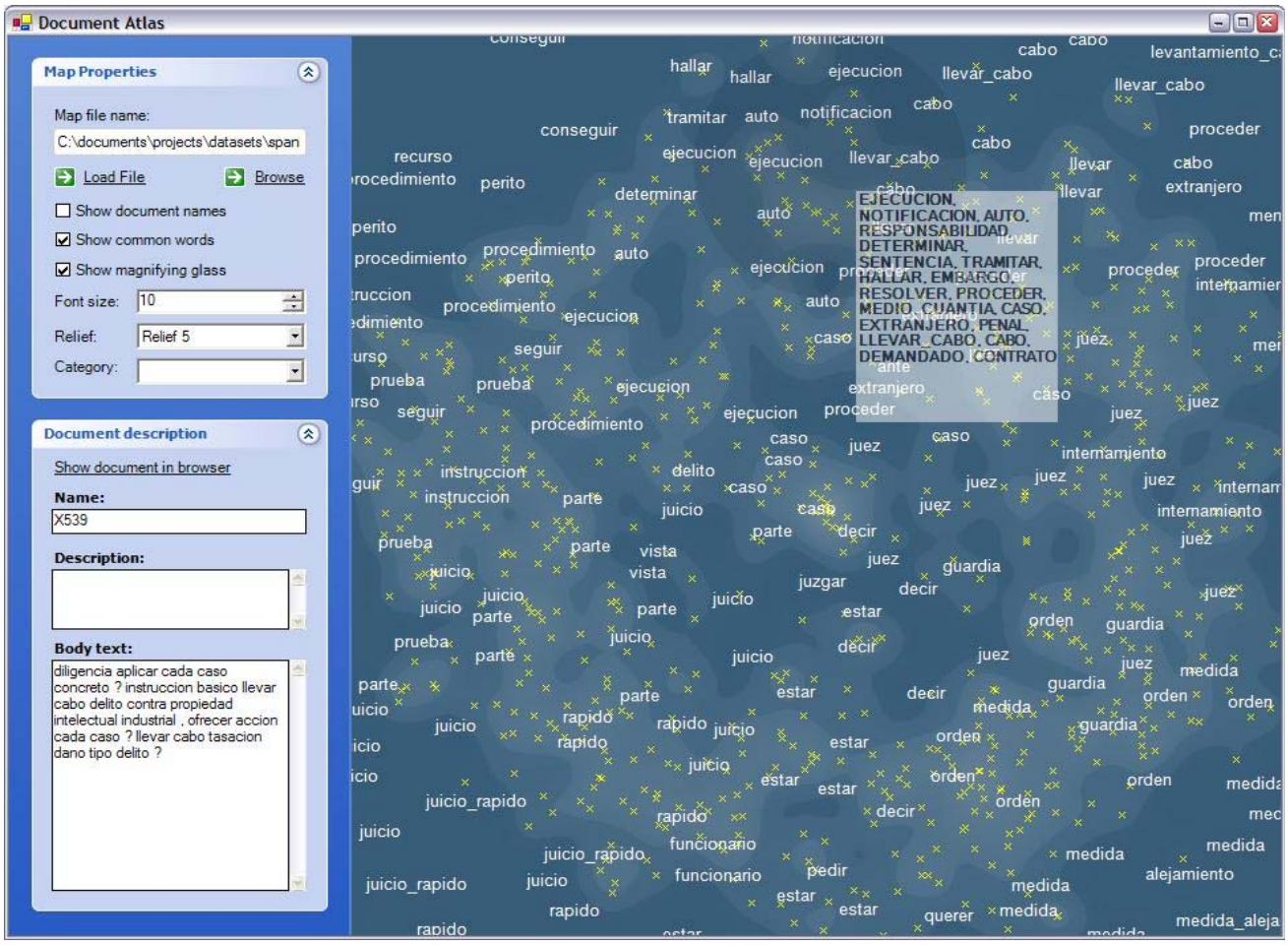

Figure 2 . Visualization of questions from Spanish judges.

We have proposed a method for efficient visualization of large data collections and describe the developed system implementing the method. The system was already successfully used for visualizing *different* kinds of document corpora – from project descriptions, scientific articles to questions and even clients of an Internet grocery store. We found that the system is very helpful for data analysis because it offers quick insight into the structure of the visualized corpus. 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. Now we use it to show the density of documents but it can also serve for showing some other attributes.

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

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

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

[4] M. Grobelnik, D. Mladenic. Analysis of a database of research projects using text mining and link analysis. In: *Data mining and decision support : integration and collaboration*, The Kluwer international series in engineering and computer science, SECS 745. Boston; Dordrecht; London: Kluwer Academic Publishers, 2003, pp. 157-166).

[5] V.R. Benjamins, J. Contreras, P. Casanovas, M. Ayuso, M. Becue, L. Lemus, C. Urios. Ontologies of Professional Legal Knowledge as the Basis for Intelligent IT Support for Judges, Workshop on *Legal Ontologies and Web-based Legal Information Management*, held at ICAIL 2003, Edinburgh, 2003.