Learning Term Spaces based on Visual Feedback

Michael Granitzer
Department of Knowledge Discovery
Know-Center Graz
Inffeldgasse 21a, 8010, Graz, Austria
mgrani@know-center.at

Thomas Neidhart
Institute of Knowledge Management
Graz University of Technology
Inffeldgasse 21a, 8010, Graz, Austria
tneidhart@know-center.at

Mathias Lux
Institute of Knowledge Management
Graz University of Technology
Inffeldgasse 21a, 8010, Graz, Austria
mlux@know-center.at

Abstract

Extracting and visualizing concepts and relationship between text documents strongly depends on the used similarity measure. In order to provide meaningful visualizations and to extract useful knowledge from document collections, user needs must be captured by the internal representation of documents, and the used similarity measure. In most applications the Vector Space Model and the Cosine similarity are used therefore and serve as good approximations. Nevertheless, influencing similarities between documents is rather hard, since parameter tuning relies heavily on expert knowledge of the underlying algorithms, and the influence of different weighting schemes and similarity measures is not known before.

In this paper we present an approach on how to adapt the vector space representation of documents by giving visual feedback to the system. Our approach starts by clustering a corpus of text documents and visualizing the results using multi dimensional scaling techniques. Afterwards, a 2D landscape visualization is shown which can be manipulated by the user. Based on these manipulations the high dimensional representation of the documents is adapted to fit the users need more precisely. Our experiments show that iterating these steps results in an adapted representation of documents and similarities, generating layouts as intended by the user and furthermore increases clustering accuracy. While this paper only investigates the influence on clustering and visualization, the method itself may also be used for increasing classification and retrieval performance since it adapts to the users need of similarity.

1. Introduction

Extracting and visualizing relationships between text documents is important for several applications in the domain of intelligence services, trend analysis and navigation in large text corpora. Systems like SPIRE, VxInsight, WEBSOM, and InfoSky (see [5], [3]) usually visualize relationships between documents by arranging them in a 2D or 3D layout based on the similarities between documents, which allows users to get an overview of a text collection and to discover previously unknown relationships. Nevertheless, the definition of similarity between documents, as well as the form of representation of a document is crucial for these approaches. The problem in defining a ”good” similarity measure is, that the similarity depends strongly on the user and his/her current task. Similarities may be different if someone wants to discover relationships between persons or between topics and therefore some features of a document (e.g. person names) may be more important than others (e.g. research topics addressed in a paper).

One well known model for formalizing the content of a document is the so called Basic Vector Space Model, introduced by Salton et. al. 1975 [6]. Thereby documents are decomposed in terms resp. words, while each term defines a dimension in the so called document term vector and the weight assigned to the term corresponds to the importance of the term in this document. All document term vectors span the corresponding vector space. This simple model achieves surprisingly good results, not only in information retrieval but also for other application areas like for example text clustering and visualization. Within the vector space model, similarities between documents are in general calculated using the inner product between their term vectors,

which corresponds to the angle between two documents if the vectors are normalized to 1. Thereby, the Vector Space Model treats terms as independent from each other, which is in general not true for natural language. The generalized vector space model (GSVM), introduced by Wang et. al. [8], addresses the problem of having correlated terms, by looking at the co-occurrence of terms. Two terms become semantically related if they significantly often co-occur within documents. Again, similarities are estimated by the inner product of two vectors, but in different to the BVSM the GSVM applies a linear transformation based on the co-occurrence information of the document vectors before similarities are calculated. Applying a linear transformation to a vector space results in a new vector space, which takes co-occurrence information into account. A similar approach is taken by spectral retrieval methods like Latent Semantic Indexing (LSI) [1], which projects the original vector space into a lower dimensional semantic concept space based on co-occurrence information between terms. Again, similarities are calculated using inner product in the transformed space.

Common to all these approaches is, that they rely on a statistically calculated document model which can not be modified easily. While similarities based on inner product are sufficient for most applications, the representation of the vector space is crucial for finding the needed relationships. This has been shown by Van Rijsbergen, who has provided a unifying theoretical framework [7] using inner product vector spaces (i.e. a Hilbert Space) for adapting similarities fitting users intuition about similarity.

Motivated by this, we present an approach to learn high dimensional representation of documents based on user defined similarities between documents. We start by clustering a document collection according to the ordinary vector space model and by mapping the collection into a low dimensional space, which is visualized as a distance preserving thematic landscape. The user interacts with this visualization for manipulating the similarity between documents which is the input to our learning algorithm. The newly obtained similarities between documents are used for transforming the high dimensional space, while preserving the user defined similarity. Tests on an artificial collection, as well as on a subset of the Reuters 21578, will show that the accuracy of the clustering process can be increased significantly and the term space reflects the notion of the user.

2 Transforming the High Dimensional Space

Within this section we will formalize our approach and introduce the basic notations used in this paper. As mentioned above, we start by assuming that a document is given as column vector $\vec{d}^T = \{w_1, w_2, \ldots, w_n\}$, where w_1 defines the weight of term t_1 in the document. If a document does not contain a term, we assume that the corresponding weight is set to 0. Weights may be determined by an arbitrary weighting scheme (see [6]) and normalized to 1. All document vectors are represented in form of the term document matrix $\mathbf{D}_{n\times m} = \{\vec{d_1}, \ldots, \vec{d_m}\}$, with n as the number of terms and m as the number of documents). Similarities are calculated using the inner product between two vectors, formally written as $\sigma_{i,j} = \vec{d_i}^T \cdot \vec{d_j}$, which leads to $\mathbf{S} = \mathbf{D}^T\mathbf{D}$ for calculating the inter document similarity matrix $\mathbf{S}$ as self correlation matrix of the term document matrix.

For transforming the vector space spanned by the term document matrix $\mathbf{D}$ we introduce a linear operator in form of a matrix $\mathbf{T}_{n\times n}$. A transformed document vector $\vec{d_i'}$ is calculated by multiplying the original vector $\vec{d_i}$ with the transformation matrix formally written as

$$\vec{d_i'} = \mathbf{T}\vec{d_i}$$

The transformation matrix $\mathbf{T}$ can be viewed as a linear function transforming one space into another, while preserving the vector space operations (see [7]). Applying $\mathbf{T}$ to the term-document matrix results in a new term-document matrix embedded in the transformed space, formally written as

$$\mathbf{D}' = \mathbf{TD}$$

and the corresponding similarity matrix

$$\mathbf{S}' = (\mathbf{TD})^T\mathbf{TD} = \mathbf{D}^T\mathbf{T}^T\mathbf{TD} = \mathbf{D\Theta D}$$

We are following the formal model as outlined in [7] where linear operators are reduced to a Hermitian matrix (i.e. symmetric and positive semidefinite), which allows us to write $\mathbf{\Theta} = \mathbf{T}^2$. We now assume that the new similarity matrix $\mathbf{S}'$ is given a priori and that we seek the linear transformation $\mathbf{\Theta}$. For doing so, we resolve the above matrix multiplication for each cell of the similarity matrix $\mathbf{S}'$ which leads to

$$\sigma'_{i,j} = \sum_{l=1}^{n}\sum_{k=1}^{n} d_{j,l} \cdot t_{l,k}^2 \cdot d_{i,k}$$

We can now rewrite the above equation in a way such that

$$\begin{aligned} a_{i*j+j,k*l+l} &= d_{j,l} \cdot d_{i,k} && \forall i = 1 \ldots m, j \geq i \wedge \\ & && \forall l = 1 \ldots n, k \geq l \\ y_{i*j+j} &= \sigma_{i,j} && \forall i = 1 \ldots m, j \geq i \\ \beta_{k*l+l} &= t_{k,l}^2 && \forall l = 1 \ldots n, k \geq l \end{aligned}$$

and

$$\mathbf{A} = \begin{pmatrix} a_{1,1} & \ldots & a_{1,n^2/2-1} \\ \vdots & \ddots & \vdots \\ a_{m^2/2-1,1} & \ldots & a_{m^2/2-1,n^2/2-1} \end{pmatrix}$$

and formulate the linear regression problem as

$$\arg\min_{\vec{\beta}} \left\{ \mathbf{A} \cdot \vec{\beta} - \vec{y} = 0 \right\}$$

Hence matrix $\mathbf{A}$ contains every dimension of the documents responsible in calculating the inner product, vector $\vec{y}$ contains the similarity values and β contains the model parameter resp. the space transformation.

3 Implementation

As described above, we are interested to visualize the relationship between documents and to allow the adaption of the relationship between documents. Applying the above transformation within these setting allows us to interactively adapt relationships between terms and documents, so that the high dimensional representation expresses the users intuition about what is similar and what is not. For doing so, we perform the following steps:

1. Initialize $\mathbf{\Theta}$ as identity matrix[1]
2. Calculate high dimensional similarity matrix considering the current transformation matrix($\mathbf{S}_{high} = \mathbf{D\Theta D}$)
3. Cluster documents based on the current similarity matrix $\mathbf{S}_{high}$
4. Apply a force directed placement (FDP) algorithm on the documents and the obtained clusters to map them into a low dimensional (i.e. 2D) space. Similarities for FDP are calculated using the current transformation matrix(see [5] for details on the FDP algorithm).
5. Visualize the low dimensional space and allow user to update the visualization for changing the low dimensional representation. User interactions include moving clusters and documents in the 2D layout
6. After user manipulation is finished, the new similarities are calculated from the 2D layout. The 2D vectors for each document are normalized (using the euclidean norm) and the low dimensional similarity matrix is $\mathbf{S}_{low}$ is calculated using the inner product of the 2D vectors of each document. Thus, if the user has moved documents closer to each other, the similarity between them is increased and decreased if the distance between documents is increased.
7. Calculate the new transformation matrix $\mathbf{\Theta}$ via linear regression solving $\mathbf{S}_{low} = \mathbf{D\Theta D}$.
8. go to 2 with the new transformation matrix $\mathbf{\Theta}$

[1] By doing so we obtain the original similarity matrix in the first step

In our prototype, interactions are restricted to moving documents and clusters. While more complex interactions like splitting and merging clusters may be possible, we will see in the experimental section below that those simple interaction concepts are sufficient for improving clustering and visualization and we expect that with more complex interaction methods results can be further improved. Additionally, our goal was not to find new interaction concepts for optimizing the visualization but rather to transform the high dimensional space based on user feedback. Furthermore, we must be aware that information is lost by transforming the high dimensional space into a 2D space. Nevertheless, different groups have shown, that important relationships are still preserved by using FDP as mapping. Furthermore our experiments indicate that high dimensional relationships can be improved despite the information loss.

4 Experiments and Results

In order to evaluate our approach we have performed experiments using a subset of the Reuters-21578 and an artificial data set. The artificial data set consits of 27 documents and 5 classes and the Reuters-21578 consisted of 42 documents and 8 classes[2].

For regression we have used an implementation of the Widrow Hoff gradient descent algorithm (see [2]). In order to keep the computations feasible we have restricted the transformation matrix $\mathbf{\Theta}$ to a diagonal matrix, which has the impact that the feedback process ignores relationship between terms and only scales dimensions of the original term space. [3]

For both data sets, we started with clustering and measured the F_1 value (see [6]) w.r.t. to the original classes. Afterwards visualization and user manipulation took place iteratively and the F_1 value for clustering in the transformed space was measured and compared to the original F_1 value.

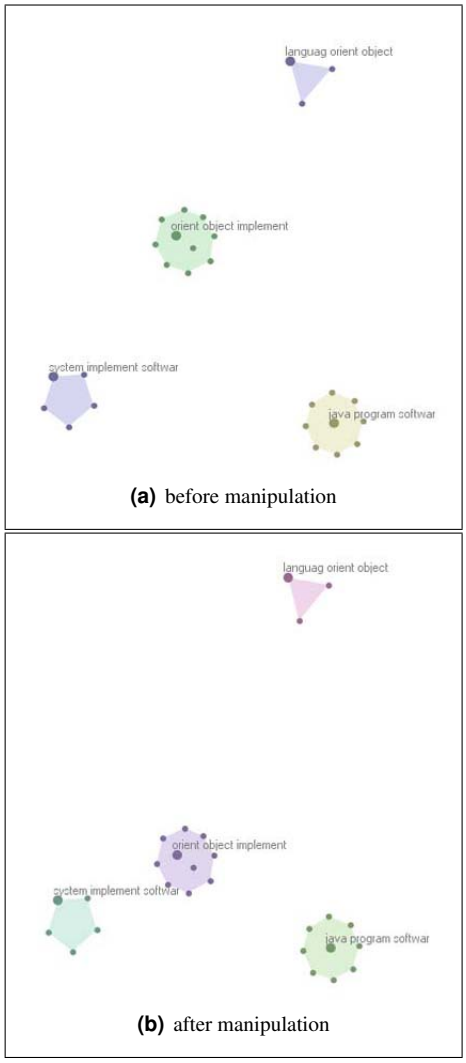
4.1 Artificial Data Set

The first evaluation was done using an artificially constructed dataset, consisting of 5 classes with 27 documents. The classes were chosen to share some terms in common, to represent a degree of similarity between them. The test evaluation is outlined in Figure 1, the first figure shows the initial state after clustering and visualization of the dataset. The second figure displays the state after user manipulations have taken place. The intent of the manipulation is to

[2] fuel, crop, farm, credit, nickel, orange, tea, coffee

[3] While this seems to be a strong constraint, experiments show that results can be improved even so. Preliminary experiments on using a full transformation matrix promise very good results, but are not yet complete for publishing.

reflect a higher degree of similarity between certain clusters, therefore the terms should be weighted according to this manipulation.



(a) before manipulation

(b) after manipulation

Figure 1. Artificial dataset: user manipulation

After calculating the transformation matrix $\mathbf{\Theta}$, we transformed the term document matrix $\mathbf{D}$ using $\mathbf{\Theta}$. The resulting term vector for a document in the moved cluster is shown in Table 1. As it can be seen certain terms have been weighted higher than before, thus reflecting a higher degree of similarity with the two clusters on the bottom side. The stemmed terms *object* and *orient* got a lower weight, because the similarity with the top cluster (which shares the same terms) has been reduced by the user due to increasing distance to the cluster.

An evaluation of the F_1 value was not necessary, as the initial state of clustering already generated a value of 1.0

Term	Weight
object	0.575
orient	0.575
implement	0.486
develop	0.226
softwar	0.226

Term	Weight
implement	0.564
object	0.485
orient	0.485
develop	0.339
softwar	0.300

Table 1. Artificial dataset: term vectors

and therefore could not be improved.

4.2 Reuters-21578 Data Set

The second evaluation was done using a subset of the Reuters-21578 data set. For this paper, we will only present a small portion of the evaluation. As it can be seen in Figure 2(a) we have chosen the cluster "fuel" as a demonstration object for our approach. The goal is to create a more compact cluster using user manipulation to reflect different document similarities. The resulting cluster can be seen in Figure 2(b)

Table 2 shows the combined term vector of the cluster "fuel". As it can be seen, the weights of the terms have been changed to reflect the user manipulations. The table only shows the 6 most significant terms, and the relevance ordering of the terms has been changed because of the transformation. Therefore the concept of the cluster has slightly been changed through user manipulation and ranks the term "fuel" (which accurately describes the cluster) top.

Term	Weight
ct	2.274
fuel	1.637
oil	1.411
number	1.375
energi	1.289
dlr	0.863

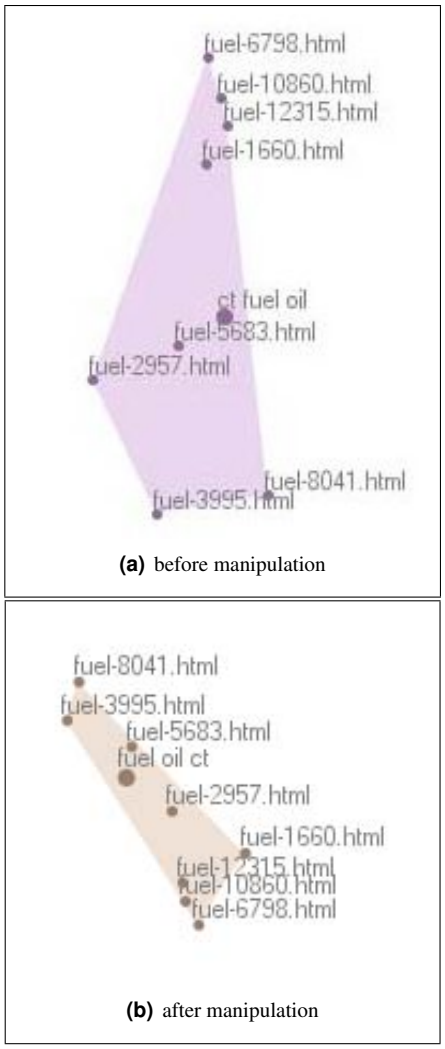
Term	Weight
fuel	3.045
oil	2.398
ct	2.129
energi	1.992
number	1.032
dlr	0.992

Table 2. Reuters dataset: term vectors

The whole evaluation consisted of several tasks, which resulted in an increase of the F_1 value from initially 0.588 to a value of 0.758.

5 Discussion and Conclusion

In our approach we have shown how high dimensional representations can be adapted towards a more accurate reflection of the users intuition of what is similar and what not. Text mining tasks like clustering and multi dimensional scaling have been improved by our approach using simple



(a) before manipulation

(b) after manipulation

Figure 2. Reuters dataset: user manipulation

interaction concepts. To reduce the complexity of the regression task, the transformation matrix was reduced to a diagonal matrix, which does not overcome the problem of the correlated terms in the vector space model. Nevertheless, by mapping the vector space into a concept space (e.g. using LSI), this drawback may be overcome. Furthermore, a full transformation matrix may be considered, whereby again techniques for reducing dimensionality resp. the degrees of freedom of the regression task have to be considered, which is planned for future work. Also, user feedback based on visualization is a limited method of including the current user context since the semantic of operations in the low dimensional space is not clearly defined. Nevertheless, since our approach only needs a suitable target similarity matrix, we can adapt our approach to other feedback methods containing more semantic information about document similarities. Taxonomies and semantically rich metadata based on conceptual graphs have the potential of deriving good similarities (see [4]) for our approach and will be considered in futher work.

Acknowledgment

These results have been developed in the MISTRAL project financed by the Austrian Research Promotion Agency (www.ffg.at) within the strategic objective FIT-IT under the project contract number 809264/9338. The Know-Center is a Competence Center funded within the Austrian Competence Centers Program K plus under the auspices of the Austrian Ministry of Transport, Innovation and Technology (www.kplus.at) and the country of Styria.

References

[1] H. Bast and D. Majumdar. Why spectral retrieval works. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 11–18, New York, NY, USA, 2005. ACM Press. 2

[2] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2001. 3

[3] M. Granitzer, W. Kienreich, V. Sabol, K. Andrews, and W. Klieber. Evaluating a system for interactive exploration of large, hierarchically structured document repositories. In *IEEE Symposium on Information Visualization, 2004. INFOVIS 2004*. IEEE, 2004. 1

[4] M. Lux and M. Granitzer. A fast and simple path index based retrieval approach for graph based semantic descriptions. In *Fachberichte Informatik, Universitt Koblenz, ISSN 1860-4471, Koblenz, Germany*, pages 29–44, 2005. 5

[5] V. Sabol, W. Kienreich, M. Granitzer, J. Becker, K. Tochtermann, and K. Andrews. Applications of a lightweight, web-based retrieval, clustering, and visualisation framework. In *Lecture Notes in Computer Science*. Springer, 2002. 1, 3

[6] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an International Journal*, 24(5):513–523, 1988. 1, 2, 3

[7] C. J. van Rijsbergen. *The Geometry of Information Retrieval*. Cambridge University Press, New York, NY, USA, 2004. 2

[8] S. K. M. Wong, W. Ziarko, and P. C. N. Wong. Generalized vector spaces model in information retrieval. pages 18–25, 1985. 2