New asymmetric iterative scaling models for the generation of textual word maps

Alberto Muñoz[1], Manuel Martin-Merino[2]

[1] Universidad Carlos III de Madrid – C/Madrid 126 – 28903 Getafe – Spain – albmun@est-econ.uc3m.es

[2] Universidad Pontificia de Salamanca – C/Compañia 3 – 37002 Salamanca – Spain – mmerino@ieee.org

Abstract

The iterative spring model (Kopcsa and Schiebel, 1998) is a kind of multidimensional scaling algorithm (MDS) based on point mass mechanics, that embeds objects in a two dimensional Euclidean space and allows to visualize object relationships and cluster structure. This technique assumes that the similarity matrix for the data set under consideration is symmetric. However there are many interesting problems where asymmetric proximities arise, like text mining problems. In this work we propose a variety of improvements to this algorithm to deal with asymmetric dissimilarities. Clustering quality and distances preservation of the resulting word maps are evaluated through objective measures. The new asymmetric algorithms outperform both, their symmetric counterpart and other widely used multidimensional scaling methods according to the objective measures computed.

Keywords: Multidimensional Scaling, Asymmetric proximities, text mining

1. Introduction

Let X be an $n \times m$ (transposed) document matrix representing n words by m documents in $\mathbb{R}^m$. Let S be the $n \times n$ matrix made up of word similarities (using a given similarity measure). We are interested in the case where the matrix S is asymmetric, that is, $s_{ij} \neq s_{ji}$. This case has been considered in the past; see for instance (Zielman and Heiser, 1996; Chen et al., 1996). When distances are used instead similarities, $d_{ij} = d_{ji}$ is assumed. To understand the need for asymmetric similarity measures, consider that word associations are not symmetric: for instance, most people will relate 'data' to 'mining' more strongly than conversely. Any similarity measure to model the relation between these two words should not obey the constraint $s_{ij} = s_{ji}$.

There are a variety of techniques suitable to generate visual representations of such word relations (word maps), such as MDS algorithms (Cox and Cox, 2001), correspondence analysis (Lebart et al., 1984; Benzécri, 1992) and neural net based algorithms (Kohonen et al., 2000; Muñoz, 1998). Word maps represent words (usually codified as vectors in a high dimensional space) as points of a two dimensional Euclidean space. Several articles have empirically shown that word maps are useful tools to discover vocabulary related to a given topic and are also valuable to model relations among different topics in databases (Chen et al., 1998; Lin, 1997)

The previously mentioned algorithms can be applied in the asymmetric case if the similarity matrix is first symmetrized (substitute s_{ij} by s^*_{ij}, where $s^*_{ij} = \frac{s_{ij} + s_{ji}}{2}$). However, information

provided by asymmetry is lost.

A number of MDS algorithms and neural net algorithms have been developed in the past (Constantine and Gower, 1978; Okada, 1997; Zielman and Heiser, 1993; Chen et al., 1996; Saito, 1991) that deal with asymmetry from different viewpoints. For the task of word map generation we are interested in algorithms able to achieve a balance between word clusters separation and distances preservation. In (Kopcsa and Schiebel, 1998) it has been presented an iterative model based on classic mechanics that complies with these two requirements. In addition, the authors claim that convergence is faster than in most iterative MDS algorithms.

In this paper we generalize the algorithm of (Kopcsa and Schiebel, 1998) in a variety of ways to the asymmetric case. The generalization will be achieved by first defining new asymmetry coefficients that convey the information provided by asymmetry, and then incorporating them into the algorithm in an appropriate manner.

The paper is organized as follows. Section 2 introduces the asymmetry coefficients. In section 3 we present the new asymmetric models. In section 4 we study the performance of our algorithms on a real text database and, finally, section 5 gets conclusions and points out some directions for future work.

2. Asymmetry

Symmetric measures have been widely used in the context of information retrieval (Rorvig, 1999) These measures fail to accurately model similarity between words in the sense that semantically close words often have low similarity coefficients (see (Muñoz, 1997) for details). In a few words, the L_1 norm of a word is the number of documents indexed by the word, and due to Zipf's law, the distribution of L_1 word norms (and therefore that of L_2 norms) is very asymmetric (see figure 1) . This fact distorts distance comparisons between words that have large differences in their norms.

In this section we first introduce two commonly used asymmetric measures that do not suffer from this drawback and next we review some coefficients of asymmetry needed for the algorithms proposed in section 3.

2.1. Asymmetry similarity measures

The first similarity measure introduced in this section, fuzzy logic similarity, has been widely used in the context of fuzzy logic models (Klir and Yuan, 1995) and information retrieval (Rorvig, 1999). The second one is a well known probabilistic measure, the Kullback-Leibler (K-L) divergence.

1. Fuzzy logic similarity is defined as

$$s_{ij} = \frac{|x_i \wedge x_j|}{|x_i|} = \frac{\sum_k |min(x_{ik}, x_{jk})|}{\sum_k |x_{ik}|}$$

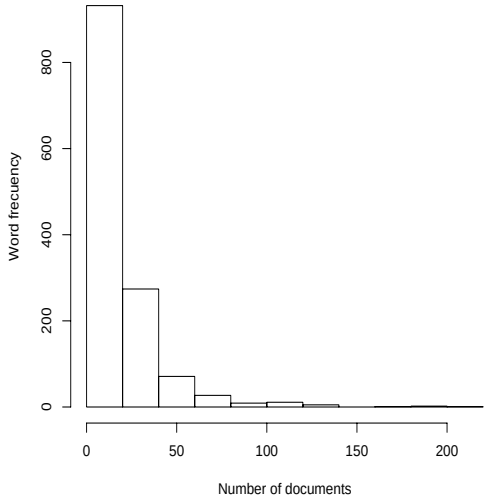

Figure 1: *Frequency histogram for terms in the database used in section 4. Number of documents is the number of documents containing a given word*

where $\wedge$ is the min fuzzy operator and L_1 norm is used. s_{ij} may be interpreted as the degree with which topic i is a subset of topic j (Kosko, 1991). Obviously $s_{ij} \neq s_{ji}$ allowing to model accurately asymmetric relationships.

2. K-L divergence (Dagan et al., 1999) is defined as

$$d_{ij} = D(x_i \| x_j) = \sum_{x_k} p(x_k|x_i) log \left(\frac{p(x_k|x_i)}{p(x_k|x_j)} \right)$$

where $p(x_k|x_i)$ denotes the probability that word x_k appears together with x_i. K-L divergence between word x_i and x_j measures the distance between the probability distribution functions associated to the context of x_i and x_j. This means that d_{ij} is small when words that appear close to x_i appear also close to x_j.

2.2. Asymmetry coefficients

Asymmetry coefficients convey the information provided by asymmetry. We define two coefficients, the first one derived from the fuzzy logic similarity introduced in section 2.1 and the second one associated to the K-L divergence defined in the same section. Let $D = (\delta_{ij})$ be the matrix of dissimilarities between object pairs. It is well known that any square non-symmetric matrix can be decomposed into a symmetric and skew-symmetric component (Zielman and Heiser, 1996) $D = S + A$ where $s_{ij} = \left(\frac{\delta_{ij}+\delta_{ji}}{2}\right)$ and $a_{ij} = \left(\frac{\delta_{ij}-\delta_{ji}}{2}\right)$. It has been shown in (Martin-Merino and Muñoz, 2001) that only the skew-symmetric component provide information about asymmetry.
If we compute a_{ij} for the fuzzy logic similarity,

$$a_{ij} \propto \frac{|x_i \wedge x_j|}{|x_i|} - \frac{|x_j \wedge x_i|}{|x_j|} \propto |x_i| - |x_j|$$

that suggests that asymmetry is due to differences in L_1 norm. So, the first coefficient we propose to model the skew-symmetric component of the dissimilarity matrix is a normalized L_1 norm $\left(|x_i| = \frac{|x_i|}{\max_l(|x_l|)}\right)$ that takes values in the $[0, 1]$ interval. Intuitively speaking, this coefficient will become large for general (broad sense) terms with large L_1 norms.

The second coefficient is related to the K-L divergence and may be computed as follows. We first transform the K-L divergence into a similarity $k_{ij} = 1 - \frac{d_{ij}}{\max d_{ij}}$ so that $k_{ij} \in [0, 1]$. Then we define the K-L coefficient of asymmetry as $k_i = \sum_k k_{ki}$, that inform about the grade with which context of x_k $(k = 1 \ldots N)$ are a subset of x_i context. Intuitively speaking this coefficient will become large for terms with a wide context (broad sense terms).

3. Iterative asymmetric spring models

In this section we briefly describe the basic idea of the iterative spring model proposed in (Kopcsa and Schiebel, 1998). Next we propose variants of the basic model that take into account the asymmetric component of the dissimilarity matrix through the asymmetry coefficients defined in section 2.2.

3.1. Iterative symmetric spring model revisited

This model is based on mass point mechanics. Each word is represented by a mass point and they are connected to each other by springs of elasticity proportional to the similarity between the words . The mass point coordinates are updated until convergence where the final point distances represent the dissimilarity between the words. Let m_i be the mass of each point, $\Delta x_{ij} = x_j - x_i$, k_i the frictional resistance coefficient and e_{ij} the elasticity coefficient. The authors take e_{ij} proportional to the transformed Jaccard similarity. Then the forces applied over particle i are

$$f_{mi} = -m_i \ddot{x}_i$$
$$f_{ki} = -k_i \dot{x}_i$$
$$f_{eij} = e_{ij} \Delta x_{ij}$$

where f_{mi} is the inertia force, f_{ki} is the frictional resistance force and f_{eij} is the elasticity force. In equilibrium the sum of all forces over particle i has to be 0, so

$$f_{mi} + f_{ki} + \sum_j f_{eij} = 0$$

or

$$-m_i \ddot{x}_i - k_i \dot{x}_i + \sum_j e_{ij} \Delta x_{ij} = 0$$

that after approximations explained in (Kopcsa and Schiebel, 1998) gives a simple iterative solution for the differential equation

$$x_i(t+1) = x_i(t) + \frac{\Delta t}{k_i} \sum_j e_{ij} \Delta x_{ij}$$

where

$$e_{ij} = \frac{J_{ij} - T}{\max_{ij}(J_{ij}) - T}$$

T is a experimental parameter that controls that forces between particles of the same cluster are attractive and forces between particles of different clusters are repulsive. This feature decreases clustering overlapping in the final word map. Δt is the step length and $k_i = 1$ for all particles.

3.2. New iterative asymmetric models

3.2.1. Incorporating asymmetry through resistance coefficients

In order to get a deeper understanding of our asymmetric models, let first study the interaction between two terms (i, j) where i is a general term and j a specific term. According to the symmetric model proposed in (Kopcsa and Schiebel, 1998) the forces f_i and f_j are given by

$$f_i = -k_i \dot{x}_i + e_{ij} \Delta x_{ij}$$

$$f_j = -k_j \dot{x}_j + e_{ij} \Delta x_{ji}$$

where $e_{ij} = e_{ji}$ and $k_i = k_j = 1$. This means that $f_i = -f_j$ and the forces over each term are the same although the specific terms should be attracted more strongly than the general terms due to asymmetry.

We propose to give each k_i a value proportional to any of the two coefficients of asymmetry defined in section 2.2. In this way, general terms will have a large resistance coefficient and specific terms a small one. Therefore, it can be easily seen that forces from general terms to specific terms f_{ij} will be stronger than forces from specific terms to general (broad sense) terms f_{ji}.

3.2.2. Incorporating asymmetry through asymmetric elasticities

Another interesting possibility to incorporate asymmetry is to define new asymmetric elasticity coefficients in the following way,

$$e_{ij} = s e_{ij}^{(s)} \frac{l_j}{max_k(l_k)}$$

where

$$se_{ij}^{(s)} = \frac{s_{ij}^{(s)} - T}{\max_{ij}(s_{ij}^{(s)}) - T}$$

T will be fixed experimentally as in the symmetric case to achieve a balance between clusters separation and distances preservation. $s_{ij}^{(s)}$ is the symmetric component of the similarity matrix s_{ij} (for instance fuzzy logic similarity) and l_j is any of the two asymmetry coefficients defined in section 2.2. Obviously the new elasticity is asymmetric and can be decomposed as

$$e_{ij} = e_{ij}^{(s)} + e_{ij}^{(a)} = \frac{se_{ij}^{(s)}}{2max_k(l_k)}(l_j + l_i) + \frac{se_{ij}^{(s)}}{2max_k(l_k)}(l_j - l_i)$$

The symmetric component would reduce to the symmetric case if all l_j are equal, that is, asymmetry does not exists. The skew-symmetric component has basically the same form of the skew-symmetric component of the similarity matrix (s_{ij}). This term increases the forces from general terms to specific terms while decreasing forces from specific terms to broad terms.

3.2.3. Incorporating asymmetry through asymmetric distances

A natural way to incorporate asymmetry and that has been widely used by the MDS community is to define the distances between the mass points as a symmetric term (Euclidean distance) plus a bias term (Okada, 1997) . Words are distributed over the map such that asymmetric distances over the final configuration approximate the word dissimilarity matrix. This fact will avoid the degradation suffered by techniques that try to embed points from a non Euclidean space into a Euclidean space. We then define vector difference between two mass points as

$$\overline{\Delta xa_{ij}} = ||\overline{\Delta xa_{ij}}||\overline{u_{ij}} = \left(||\overline{\Delta x_{ij}}|| + \frac{l_j - l_i}{2max_k(l_k)}\right)\overline{u_{ij}}$$

again, the symmetric component of $\overline{\Delta xa_{ij}}$ reduces to the Euclidean case. The skew-symmetric component allows that $d_{ji} > d_{ij}$ in the final visual map if j is a broader term than i. By substituting $\overline{\Delta x_{ij}}$ into the expression of the elastic force we get

$$f_{eij} = se_{ij}^{(s)}\left(||\overline{\Delta x_{ij}}|| + \frac{l_j - l_i}{2max_k(l_k)}\right)\overline{u_{ij}}$$

This expression shows that forces due to terms with large asymmetry coefficient get stronger due to asymmetry. Therefore general terms will become more influent in the final word map.

4. Experimental results

Assessing the performance of algorithms that produce visual word maps may not be an easy task. We will first evaluate the ability of the mapping algorithms to preserve word distances.

But from the practical point of view it is even more important to evaluate the ability of the mapping algorithms to preserve the cluster structure of the document collection. For this purpose we need a thesaurus that allows us to determine which words belong to each cluster. To this aim we have built a database made up of 982 documents that group in three main topics ('Library Science' , 'Science and Technology' and 'Economy and Sociology'). There are 7 well defined subtopics. Therefore a clustering algorithm may be run over the final word map and cluster overlapping may be easily evaluated. Mapping algorithms that preserve object dissimilarities and at the same time favor cluster separation are considered better.

Two characteristics are evaluated through well known objective measures:

1. Distance order preservation : This feature is evaluated through the Spearman rank correlation coefficient (Croft, 2000) that measures neighbors order preservation by the mapping algorithm. Note that this coefficient is weaker than the correlation coefficient and therefore more appropriate to the problem at hand. This coefficient is computed both, for 10% of the smallest distances and for all the distances. These two measures allow us to study separately the performance of the mapping algorithms for small distances and for the large ones.
2. Clustering structure preservation: To evaluate this feature, it is necessary to run a clustering algorithm (for instance PAM (Kaufman and Rousseeuw, 1990)) over the reduced dimensional space where the points are embedded by the mapping algorithms. We propose 3 measures to evaluates the cluster quality:
   - F measure (Croft, 2000) that has been widely used by the IR community. It is a compromise between "Recall" and "Precision". "Recall" gives the average maximum probability that a word of class i is assigned to a cluster j $(j = 1, 2 \dots 7)$ and "Precision" is the average maximum probability that a word of cluster j is assigned to class i $(i = 1, 2 \dots 7)$. Intuitively, F measures if words from the same class are clustered together and vice versa.
   - Entropy measure (Strehl et al., 2000): Gives the uncertainty for the classification of words from the same cluster. It achieves the maximum when the probability of points that belong to a given cluster being classified to each class is $1/g$, where g is the number of clusters. Therefore low values are considered better.
   - Mutual Information (Strehl et al., 2000): Is a nonlinear correlation measure between the word classification induced by the thesaurus and the word classification given by the clustering algorithm.

Before applying the mapping algorithms, documents are submitted to a standard text processing after which we end up with 981 documents in $\mathbb{R}^{1333}$. Words codified as vectors of 1-0 in $\mathbb{R}^{981}$ were normalized according to the L_2 norm. This preprocessing improve the performance of all techniques proposed. All algorithms were initialized by a classic MDS algorithm to avoid that any algorithm get stuck in a local minima. The only critical parameter that need to be determined for the spring model is T. T is taken for all experiments as the 0.75 quantile of the similarity matrix s_{ij} defined in section 2.1.

Experimental results are shown in table 1.

First and second columns give the Spearman rank correlation coefficient for all distances and for only the 10% of smaller distances respectively. Next three columns measure the F coefficient, the average entropy of the clusters and the mutual information. Higher coefficients are consid-

	Sp	Sp(neig)	F	Ent.	M. Inf
(1) Sym. Sammon	0.18	0.21	0.47	0.53	0.19
(2) Asym. Sammon	0.29	0.20	0.43	0.58	0.15
(3) Sym. spring	0.27	0.22	0.50	0.51	0.20
(4) $k_i = L_1$	0.29	0.22	0.52	0.50	0.21
(5) $k_i = \sqrt{L_1}$	0.28	0.22	0.53	0.50	0.21
(6) $k_i = K - L$	0.28	0.23	0.51	0.48	0.20
(7) $k_i = \sqrt{K - L}$	0.30	0.26	0.52	0.48	0.20
(8) $e_{ij} = s_{eij}^{(s)} * L_1$	0.29	0.22	0.51	0.51	0.20
(9) $e_{ij} = s_{eij}^{(s)} * K - L$	0.30	0.26	0.54	0.49	0.20
(10) Asym. dist. L_1	0.32	0.26	0.56	0.50	0.21
(11) Asym. dist. K-L	0.32	0.25	0.54	0.49	0.21

Table 1: *Comparison of spring asymmetric models and other MDS techniques*

ered better except for the entropy measure. Rows (1) and (2) refer to symmetric and asymmetric Sammon algorithms and (3) to the symmetric spring model. (4), (5), (6), (7) report results of asymmetric spring models presented in section 3.2.1 with resistance coefficients proportional to the L_1 norm, square root of L_1 norm, K-L coefficient and square root of K-L coefficient respectively . Rows (8), (9) test models presented in section 3.2.2 with asymmetric elasticities and asymmetry coefficients proportional to the L_1 norm and the K-L coefficient respectively . And finally rows (10), (11) report results on models presented in 3.2.3 with asymmetry coefficient proportional to L_1 norm and K-L coefficient respectively .

According to table 1, neighbor order preservation is better for the asymmetric proposed algorithms than for the symmetric spring model and the MDS algorithms. Distances preservation is improved for both, large and small distances.
F measure shows that asymmetric algorithms preserve better cluster structure of data than their symmetric counterpart. Notice that our algorithms are clearly superior to the MDS algorithms according to F measure. Moreover, entropy is smaller for our proposed asymmetric techniques, that means overlapping between the clusters is smaller. This can be explained by the fact that MDS function error only try to preserve dissimilarities but not data clusters. Finally Mutual Information slightly increases when asymmetry is incorporated. This can be justified by the fact that Mutual Information penalizes terms with large L_1 norm.
Table 1 shows that the best way to introduce asymmetry is by defining asymmetric distances. In this case all measures are clearly improved.

Finally we show in figures 2 and 3 the visual map generated by asymmetric Sammon algorithm and the asymmetric spring model. Each color denotes a different class according to the classification induced by the thesaurus. Note that it is easier to capture cluster structure for the second map and that overlapping is reduced.

5. Conclusions and future research trends

In this work we have proposed new versions of a class of MDS algorithms to deal with data where the dissimilarity matrix is asymmetric. The new algorithms have been tested on a challenging and interesting text mining problem. The models proposed are compared with both,

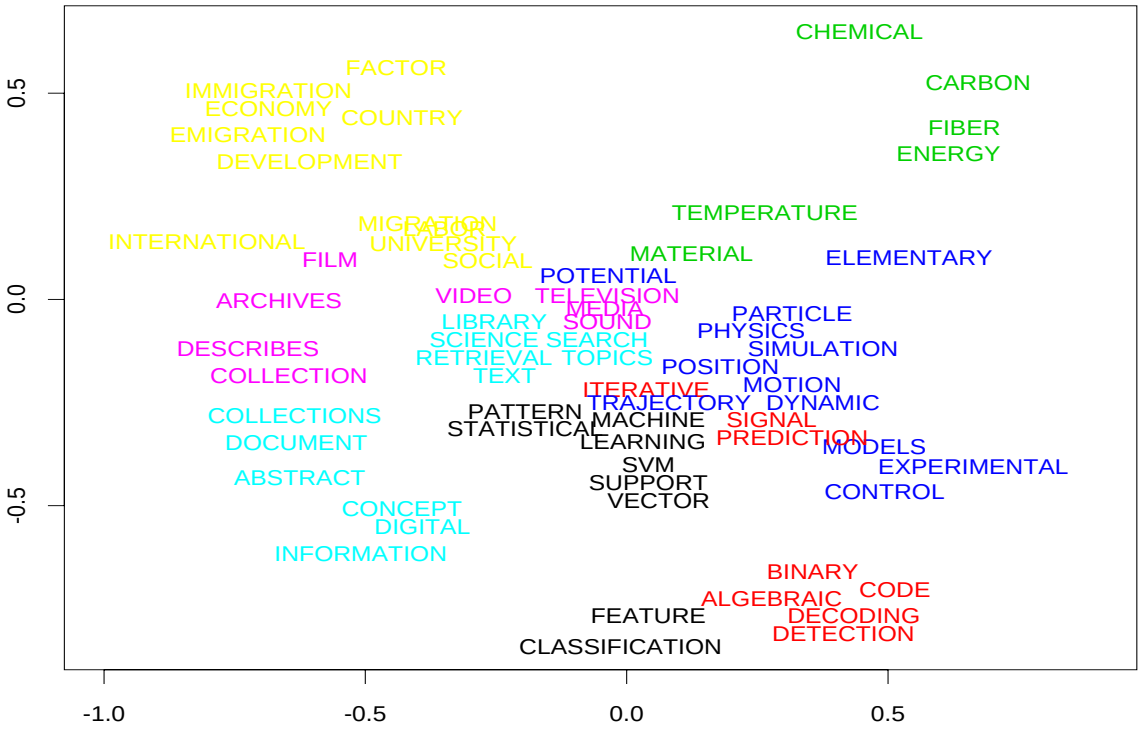

Figure 2: *Word map generated by asymmetric Sammon algorithm*

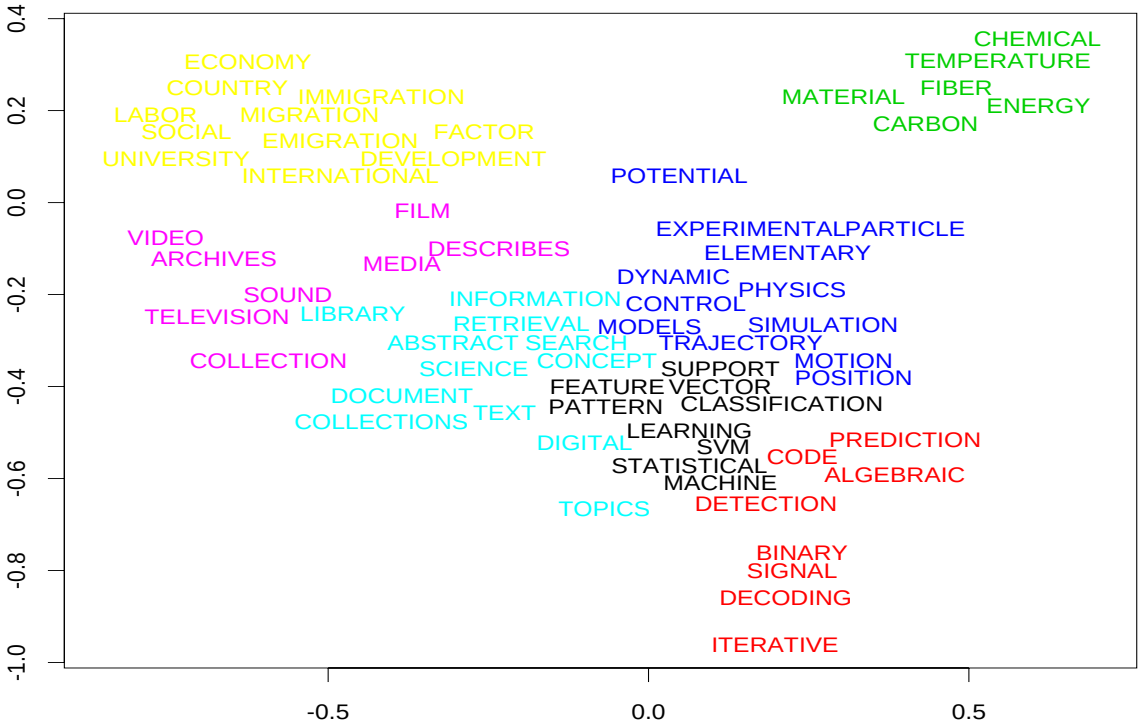

Figure 3: *Word map generated by asymmetric iterative spring algorithm*

symmetric version of the algorithms and with other iterative MDS models (also asymmetric MDS).
The asymmetric spring models improve the ability of the symmetric mapping version to separate clusters and clearly outperform the behavior of MDS algorithms for this task. Moreover, order preservation of neighbors is stronger for the proposed asymmetric versions, both for the nearest neighbors and for the last ones.
We plan in the future to study new coefficients of asymmetry derived from asymmetric measures. We will focus on the development of asymmetric hierarchical models for text mining problems.

References

Benzécri J.-P. (1992). *Correspondence Analysis Handbook*. Marcel Dekker, New York.

Chen H., Houston A. L., Sewell R. R., and Schatz B. R. (1998). Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Science*, 49(7):582–603.

Chen H., Schatz B., Ng T., Martinez J., Kirchhoff A., and Lin C. (1996). A paralell computing approach to creating engineering concept spaces for semantic retrieval: The illinois digital library initiative project. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):775–782.

Constantine A. G. and Gower J. C. (1978). Graphical representation of asymmetric matrices. *Applied Statistics*, 27(3):297–304.

Cox T. F. and Cox M. A. A. (2001). *Multidimensional Scaling*. Chapman and Hall/CRC, 2nd, USA.

Croft W. B. (2000). *Advances in Information Retrieval*. Kluwer Academic, USA.

Dagan I., Lee L., and Pereira F. C. N. (1999). Similarity-based models of word coocurrence probabilities. *Machine Learning*, 34:43–69.

Kaufman L. and Rousseeuw P. J. (1990). *Finding groups in Data; An Introduction to Cluster Analysis*. John Wiley and Sons, USA.

Klir G. J. and Yuan B. (1995). *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice Hall, USA.

Kohonen T., Kaski S., Lagus K., Salojarvi J., Honkela J., Paatero V., and Saarela A. (2000). Organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11(3):574–585.

Kopcsa A. and Schiebel E. (1998). Science and technology mapping: A new iteration model for representing multidimensional relationships. *Journal of the American Society for Information Science*, 49(1):7–17.

Kosko B. (1991). *Neural Networks and Fuzzy Systems: A Dynamical Approach to Machine Intelligence*. Prentice Hall, Englewood cliffs.

Lebart L., Morineau A., and Warwick J. F. (1984). *Multivariate Descriptive Statistical Analysis*. John Wiley, New York.

Lin X. (1997). Map displays for information retrieval. *Journal of the American Society for Information Science*, 48(1):40–54.

Martin-Merino M. and Muñoz A. (2001). Self organizing map and sammon mapping for asymmetric proximities. In *Proceedings of the International Conference on Artificial Neural Neworks, LNCS 2130*, pages 429–435. Springer Verlag.

Muñoz A. (1997). Compound key word generation from document databases using a hierarchical clustering art model. *Journal of Intelligent Data Analysis*, 1(1).

Muñoz A. (1998). Self-organizing maps for outlier detection. *Neurocomputing*, 18:33–60.

Okada A. (1997). Asymmetric multidimensional scaling of two-mode three-way proximities. *Journal of Classification*, 14:195–224.

Rorvig M. (1999). Images of similarity: A visual exploration of optimal similarity metrics and scaling properties of trec topic-document sets. *Journal of the American Society for Information Science*, 50(8):639–651.

Saito T. (1991). Analysis of asymmetry proximity matrix by a model of distance and additive terms. *Behaviormetrika*, 29:45–60.

Strehl A., Ghosh J., and Mooney R. (2000). Impact of similarity measures on web-page clustering. In *Proceedings of the 17th National Conference on Artificial Intelligence: Workshop of Artificial Intelligence for Web Search , Austin, Texas, USA*, pages 58–64. AAAI.

Zielman B. and Heiser W. J. (1993). Analysis of asymmetry by a slide-vector. *Psychometrika*, 58(1):101–114.

Zielman B. and Heiser W. J. (1996). Models for asymmetric proximities. *British Journal of Mathematical and Statistical Psychology*, 49:127–146.