# Topic Identification in Soft Clustering using PCA and ICA

Leonid Zhukov
Yahoo! Research Labs
Pasadena, CA
leonid.zhukov@overture.com

David Gleich [*]
Harvey Mudd College
Claremont,CA
dgleich@cs.hmc.edu

## ABSTRACT

Many applications can benefit from soft clustering, where each datum is assigned to multiple clusters with membership weights that sum to one. In this paper we present a comparison of principal component analysis (PCA) and independent component analysis (ICA) when used for soft clustering. We provide a short mathematical background for these methods and demonstrate their application to a sponsored links search listings dataset. We present examples of the soft clusters generated by both methods and compare the results.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering

## General Terms

Algorithms, Experimentation

## Keywords

Principal Component Analysis (PCA), Independent Component Analysis (ICA), Latent Semantic Indexing (LSI), clustering

## 1. INTRODUCTION

In soft clustering we assume that objects can belong to multiple categories. We propose to use PCA and ICA to generate the categorization by using the projection axes as the clusters. PCA has two properties that facilitate this use. First, PCA processing reduces the dimensionality of the data by finding the directions of maximum variance within the dataset [4]. These directions are the projection axes for PCA. Second, PCA "spheres" the data by scaling the variance along different directions and gives each axis, or cluster, equal weight. By construction, however, PCA is restricted to a set of orthogonal axes, i.e. uncorrelated clusters.

ICA [6], instead, can recover the intrinsic structure of the data by relaxing the orthogonality constraint [6]. Within the ICA model, the projection axes can be aligned with the data, even when non-orthogonal.

We apply these methods to data from Overture's (a Yahoo! subsidiary) sponsored links listing data which is a set of search terms with bids from various advertisers, i.e. a term-advertiser matrix.

## 2. METHOD

We employ the vector-space model [1] and consider a term-advertiser matrix $A$, where every column corresponds to an advertiser and every row to a bidded search term. Thus, every column of the matrix shows a bidding pattern for an advertiser and every row shows bids on the particular term. This arrangement is analogous to standard term-document matrix used in latent semantic indexing (LSI) literature [3]. Any bidded term, i.e row of the matrix, $t_i$ can be considered as a vector (data point) in the advertiser space.

We then use singular-value decomposition [5] of the matrix $A$ to establish an orthogonal coordinate system in the both advertiser and term spaces $A = USV^T$. The columns of $U$ form an orthogonal basis in term space, and the columns of V form an orthogonal basis in advertiser space.

After obtaining projections of the data with PCA and ICA (vide infra), we compute a set of identifying terms for each cluster, or projection axis. The terms associated with a particular axis are the terms with maximum projection on that axis [2].

### 2.1 PCA

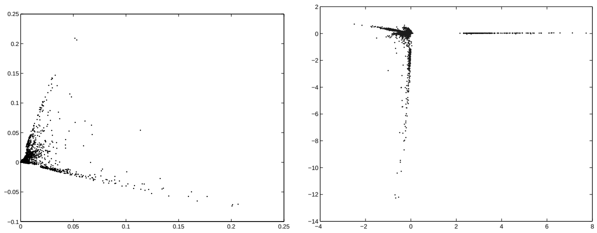
Principal component analysis uses the eigen-decomposition of the correlation matrix $M = AA^T$ to find orthogonal directions with total maximal variance of projections, $MU = \Lambda U$. PCA sorts the axis of the reduced dimensionality basis according to the total variance of the projection and retains the $k$ largest axes, thus removing redundancy and reducing the dimensionality of the data. PCA can also be calculated using the singular value decomposition of $A$ [5]. Since we are interested in term space, we will perform projection of a transposed $A^T$ matrix onto $U$ basis by $V_k^T A^T$.

$$X_k = S_k^{-1} V_k^T A^T = U_k^T$$

The covariance matrix $X_k X_k^T \approx I$ is diagonal, that is, the projection axes are uncorrelated with equal variance.[1] The

[*] Work performed while at Yahoo! Research Labs.

[1] Due to the sparsity and high dimensionality of the data the mean was already close to 0, and we chose not remove the mean from the data.



(a) Two top PCs.

(b) Two ICs.

**Figure 1: Projection of all terms onto a two-dimensional subspace formed by the top two principal components and independent components, respectively.**

matrix $X_k$ is also an optimal reduced dimensional representation of the term vectors from $A$, in the least-squared distance sense.

## 2.2 ICA

Independent component analysis finds a set of directions in the data such that when the data points are projected onto these directions, the resulting data are *statistically independent* (a much stronger condition that uncorrelated). Unlike PCA, these directions need not be orthogonal within the original space.

We employ FastICA [7] algorithm to reconstruct a matrix $W$ such that,

$$Y_k = WX_k,$$

where the rows of $Y$ are statistically independent, and $X_k$ is the reduced dimensional representation of the terms from PCA. Thus, after the PCA "sphering" procedure, an ICA algorithms only needs to adjust the axes.

## 3. DATA

In this study, we use a small, densely connected subset of Overture Services' term-advertiser data with 10,000 bidded search terms, 8,850 advertisers, and more than 250,000 bids. Before computing the PCA and ICA of the data, we normalized the rows of $A$, that is, the terms.

## 4. RESULTS

As seen from the following figures, advertisers tends to exhibit grouping behavior in their bidding patterns, i.e form clusters corresponding to market segments. These clusters stretch along preferred directions in reduced dimensional space. However, PCA enforces orthogonality of the new basis axes, and thus, in PCA space, the preferred directions might not be aligned with the PCA axes. Consequently, points might have projections on multiple axes. ICA relaxes that restriction and allows each axis to follow the clusters more precisely. In ICA space, then, points are nicely aligned along the axis and have strong projection only on one of the axis as seen in Fig.1.

*site traffic*:
18.3% cluster 182 (promotion site web, promotion web, traffic web, ...)
6.1% cluster 98 (marketing, advertising, business, ...)
3.7% cluster 199 (internet marketing, marketing online, marketing web)
2.7% cluster 4 (dating, love, single, ...)
2.3% cluster 74 (investment, investing, broker, ...)

**Table 1: The top 5 clusters for the term *site traffic*.**

| PCA axis | value | ICA axis | value |
|---|---|---|---|
| business home | 0.208 | business home | 0.379 |
| based business home | 0.197 | based business home | 0.341 |
| work home | 0.179 | business opportunity | 0.307 |
| business opportunity | 0.176 | work home | 0.276 |
| business home opport. | 0.163 | from home work | 0.249 |
| PCA axis | value | ICA axis | value |
| loss weight | 0.193 | loss weight | 0.365 |
| lose weight | 0.131 | lose weight | 0.276 |
| design site web | 0.129 | diet | 0.249 |
| diet | 0.124 | diet loss weight | 0.207 |
| design web | 0.119 | loss program weight | 0.201 |
| PCA axis | value | ICA axis | value |
| stock trading | 0.106 | stock | 0.237 |
| stock | 0.103 | stock trading | 0.236 |
| travel | 0.098 | market stock | 0.219 |
| vacation | 0.094 | investing | 0.216 |
| investing | 0.093 | investment | 0.194 |

**Table 2: Terms with maximum projections on three principal and independent components. While the projections on the first axis are clean, the projections on the 2nd and 3rd principal components show topic mixing, whereas the independent components do not.**

We can label each axis or cluster with the terms that have the largest projection on that axis. In Table 2, we see that some PCA axes have "mixed" topic terms associated with them, while the ICA axis labels are much more uniform, i.e. they are "cleaner."

Every search term, in turn, can be associated with several clusters by looking at its normalized projection values on each axis. For example, Table 1 presents the clusters for the term *site traffic* in a ICA subspace with 200 axes. Of these clusters, the first three are very relevant to the query term, while the next two are not obviously relevant. The term is more closely associated with generic site promotion than explicit marketing because of its weak membership in the "advertising" cluster.

## 5. CONCLUSIONS

We investigated applications of ICA and PCA for soft clustering and topic identification. PCA allows us to reduce the dimensionality of the data, while ICA provides superior identification of the topics.

## 6. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison Wesley, 1999.

[2] E. Bingham, A. Kaban, and M. Girolami. Topic identification in dynamical text by complexity pursuit. *Neural Processing Letters*, 17:69–83, 2003.

[3] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[4] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* John Wiley and Sons, 2001.

[5] G. H. Golub and C. F. V. Loan. *Matrix Computations.* John Hopkins Univ. Press, 1989.

[6] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

[7] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9:1483–1492, 1997.