2008

Uncovering the structure of hypergraphs through tensor decomposition: an application to folksonomy analysis

Flavian Vasile
Iowa State University

Recommended Citation

Vasile, Flavian, "Uncovering the structure of hypergraphs through tensor decomposition: an application to folksonomy analysis" (2008). *Retrospective Theses and Dissertations*. 14918.
https://lib.dr.iastate.edu/rtd/14918

Uncovering the structure of hypergraphs through tensor decomposition: An application to folksonomy analysis.

by

Flavian Vasile

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Computer Science

Program of Study Committee:
Vasant Honavar, Major Professor
David-Fernandez Baca
Samik Basu

Iowa State University

Ames, Iowa

2008

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

This thesis would not have been possible without the support of many people. The author wishes to express his gratitude to his supervisor, Prof. Dr. Vasant Honavar who was extremely understanding and supportive during the last years. Special thanks go also to all his AI lab colleagues and especially to his friend Adrian Silvescu who gave him insights and precious advice during all the graduate studies. Last but not least, he wants to thank his family for all the love and support during this period.

ABSTRACT

Folksonomies - shared vocabularies generated by users through collective annotation (tagging) of web-based content, which are formally hypergraphs connecting users, tags and objects, are beginning to play an increasingly important role in social media. Effective use of folksonomies for organizing and locating web content, discovering and organizing user communities in order to facilitate the contact and collaboration between users who share parts of their interests and attitudes calls for effective methods for discovering coherent groupings of users, objects, and tags. We empirically compare the results of several folksonomy clustering methods using tensor decompositions such as PARAFAC, Tucker3 and HOSVD which are generalizations of principal component analysis and singular value decomposition with standard methods that use 2-dimensional projections of the original 3-way relationships. Our results suggest that the proposed methods overcome some of the limitations of 2-way decomposition methods in clustering folksonomies.

CHAPTER 1. OVERVIEW

The world wide web (or simply the web) has revolutionized the way we generate, find, share, and use information resources. The success of the web can be largely attributed to the network effect: The absence of central control on content and organization of the web allows thousands of independent actors to contribute resources (e.g., web pages) that are interlinked to constitute the web. The success of this paradigm has led to the development of extremely useful resources such as the Wikipedia which now rivals several authoritative encyclopedias in terms of its coverage and accuracy Editorial (2005). Recent efforts to extend the web into a semantic web (Berners-Lee et al. (2001)) are aimed at enriching the web with machine-interpretable metadata in order to facilitate indexing, semantics-driven search (as opposed to keyword-based search) and navigation of web content. Effective use of such metadata can dramatically improve the relevance of results that search engines return in response to user-generated queries. Realizing the full potential of the semantic web requires the large-scale development, adoption and use of metadata for describing web content. The role played by the network effect in the success of the web and collaboratively generated content such as the Wikipedia strongly argues for involving the producers and users of web content in creating and associating machine-readable metadata with web content.

Folksonomies, shared vocabularies generated by users through collective and hence collaborative tagging of web-based content, constitute an important source of metadata or annotation of web content. Folksonomies can be represented as hypergraphs that link users, tags and objects. Unlike ontologies, that provide precise logical descriptions of a domain of interest in terms of the properties and relationships between objects, folksonomies lack explicit logical relationships between tags. Nevertheless, folksonomies reflect community consensus on the *latent semantics* of web content. Hence, folksonomies offer a relatively inexpensive, but powerful source of information for not only organizing web content, but also for discovering and organizing user communities (social networks), therefore helping users to locate both web content and other users who share their interests and attitudes, and facilitate collaboration.

Exploiting the full potential of folksonomies to enrich user interaction with the web calls for effective methods for discovering coherent groupings of users, objects, and tags. Consequently, there is a growing interest in methods for uncovering potentially useful regularities among relationships that link users, tags, and objects within a folksonomy (Heymann and Garcia-Molina (2006); Schmitz (2006); Begelman et al. (2006); Specia and Motta (2007)). Current approaches to analysis of folksonomies typically reduce the ternary relationships between users, tags, and objects in a folksonomy first into 2-way relationships and apply standard singular value decomposition (SVD) to the resulting 2-dimensional matrices. In a multi-way setting the reliance on projecting a higher order matrix into 2-dimensional matrices makes the results difficult to interpret. Consequently, several multi-way methods have been developed. These can be grouped in *tensor* and *hypergraph methods.* The tensor methods appeared in areas such as chemometrics and psychometrics to analyze 3-way relationships in data. Two such methods are PARAFAC (Harshman and Lundy (1994); Carroll and Chang (1970)), and Tucker3 (Tucker (1966)), originally developed in psychometrics, and subsequently applied to problems in chemometrics (Smilde (1992)). PARAFAC can be viewed as a constrained version of Tucker3, and Tucker3 a constrained version of two-way Principal Components Analysis (Kiers and Van Mechelen (2001)). Hence, any data set that can be modeled adequately with PARAFAC can thus also be modeled by Tucker3 or 2-way PCA of unfolded matrices. However, PARAFAC uses fewer degrees of freedom to model the data, and hence is attractive from the standpoint of Occam's razor. A third tensor decomposition method is HOSVD (High-Order Singular Value Decomposition) which is similar to Tucker3 but requires the vectors to be orthogonal with each other.

The other type of multi-way methods generalize graph-based clustering and ranking methods to hypergraphs.

Similarly to graphs and adjacency matrices, hypergraphs and tensors are interchangeable representations of high-dimensional relational data.

To the best of our knowledge, multi-way decomposition methods have not been used in clustering the 3-way relationships in a folksonomy. Against this background, we explore their application to clustering folksonomies and compare their results with standard PCA and spectral clustering techniques on 2-way projections of the original 3-way folksonomy matrix. We also explore the applications of folksonomy clustering in the discovery of lightweight ontologies and coherent user communities from folksonomies.

The rest of this paper is organized as follows: Section 2 introduces a tensor-based mathematical formalism for describing folksonomies, and formulates the problems of lightweight ontology extraction and discovery of coherent user communities as essentially problems of clustering 3-way relationships in a taxonomy. Section 3 briefly summarizes the SVD methods for 2-way decomposition of binary relationships represented by a 2nd order tensor (i.e., a matrix) and introduces the PARAFAC, Tucker3 and HOSVD methods for 3-way decomposition of ternary relationships represented by 3rd order tensors. Section 4 describes the application of our proposed methods to discover lightweight ontologies and coherent user communities by clustering the ternary user-tag-object relationships in a folksonomy. Section 5 describes the experimental setup and the results. Section 6 concludes with a summary discussion of the results, related work, and a outline of some directions for further research.9

CHAPTER 2. FOLKSONOMIES AND FOLKSONOMY CLUSTERING

2.1 Preliminaries.

Tensors are higher order generalizations of vectors and matrices. Informally, a tensor is represented as a multi-dimensional array relative to a choice of basis of the particular space on which it is defined. Vectors can be viewed as 1st-order tensors and matrices as 2nd-order tensors. Vectors can be used to represent unary relationships (properties) of objects; Matrices can be used to represent binary relationships between objects. Tensors of order N can be similarly used to represent N-ary relationships among objects. A tensor written in component form is an indexed array. The *order* of a tensor is the number of indices required to index the elements of the array.

A real-valued tensor of order N is denoted by $\bar{X} \in \mathcal{R}^{I_1 \times I_2 \times \ldots \times I_N}$ where $I_1 \cdots I_N$ denote the dimensions of the array representing the tensor (i.e., the number of elements in a vector, the number of rows and columns of a matrix, etc.) The dimensions of the corresponding N-dimensional array are called *modes*. Elements of $\bar{X}$ are denoted as $x_{i_1 \ldots i_n \ldots i_N}$ where $1 \leq i_n \leq I_n$. We use X to denote a matrix (tensor of order 2) and x to denote a vector (tensor of order 1).

Of particular interest to us are tensors of order 3. The first two dimensions (modes) of a 3-

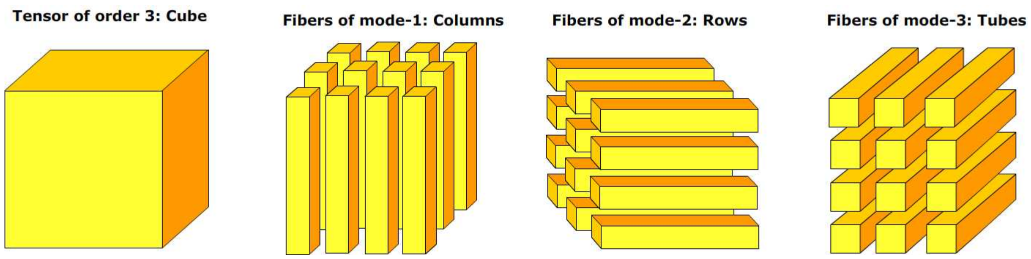

Figure 2.1 A tensor and its directions

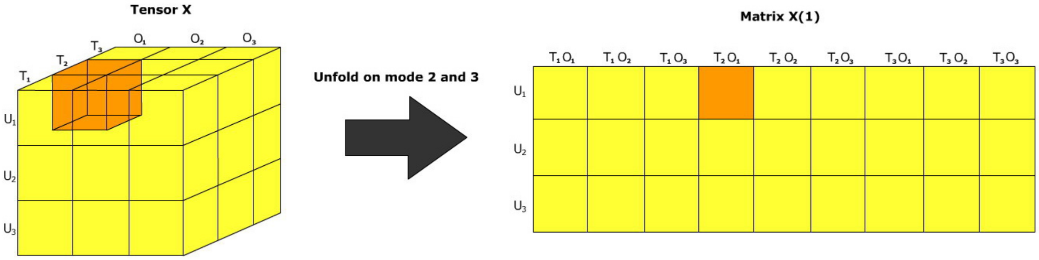

Figure 2.2 Tensor Unfolding

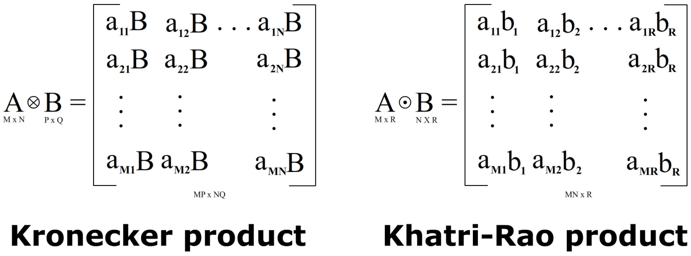
$\underset{M\times N}{A} \otimes \underset{P\times Q}{B} = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1N}B \\ a_{21}B & a_{22}B & & a_{2N}B \\ \vdots & \vdots & & \vdots \\ a_{M1}B & a_{M2}B & & a_{MN}B \end{bmatrix}_{MP\times NQ}$
$\underset{M\times R}{A} \odot \underset{N\times R}{B} = \begin{bmatrix} a_{11}b_1 & a_{12}b_2 & \dots & a_{1R}b_R \\ a_{21}b_1 & a_{22}b_2 & & a_{2R}b_R \\ \vdots & \vdots & & \vdots \\ a_{M1}b_1 & a_{M2}b_2 & & a_{MR}b_R \end{bmatrix}_{MN\times R}$
Kronecker product
Khatri-Rao product

Figure 2.3 Tensor Products

dimensional array representing a rank 3 tensor are referred to as rows and columns. The third is called a *tube* (see 2.1). *Unfolding* is the process of slicing a multidimensional array (see fig.2.2) along a chosen dimension and concatenating the resulting slices. Thus, the unfolding of $\bar{X}$ along mode n, $X_{(n)} \in \mathcal{R}^{I_n \times I_1 I_2 \cdots I_{n-1} I_{n+1} I_{n+2} \cdots I_N}$ contains the element $x_{i_1 \dots i_n \dots i_N}$ at row number $(i_{n+1}-1)I_1 I_2 I_{n-1} I_{n+2} \cdots I_N + \cdots + (i_N - 1) I_1 I_2 I_3 I_{n-1} + \cdots + (i_1 - 1) I_2 I_3 \cdots I_{n-1} + \cdots i_{n-1}$. Unfolding of a 3rd order tensor is illustrated in Figure 2.2.

The Kronecker product of two matrices A and B is designated by $A \otimes B$. When A is an $m \times n$ matrix, and B is a $p \times q$ matrix, $A \otimes B$ is an $mp \times nq$ matrix of elements obtained by pairwise multiplying the elements of the two matrices. Kronecker product is a special case of tensor product. (fig.2.3)

The Khatri-Rao product of two matrices A and B is designated by $A \odot B$. If A is an $m \times r$ matrix, and B is a $n \times r$ matrix, then $A \odot B$ is an $mn \times r$ matrix of elements obtained column-wise multiplication of the elements of the two matrices. Khatri-Rao product is a special case of Kronecker product. (fig.2.3)

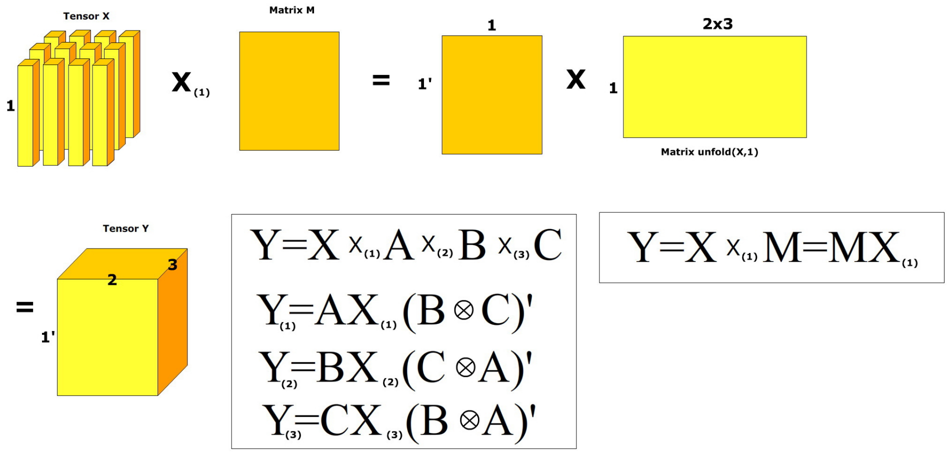

Figure 2.4 Tensor times matrix

The mode-n product of a tensor $\bar{X} \in R^{I_1 \times I_2 \ldots \times I_{n-1} \times I_n \times I_{n+1} \times \ldots \times I_N}$ with a matrix $M \in R^{J_n \times I_n}$ is denoted by $X \times_n M$ and results in a tensor $\bar{Y} \in R^{I_1 \times I_2 \times \ldots \times I_{n-1} \times J_n \times I_{n+1} \times \ldots \times I_N}$ whose entries are computed as the following: $(X \times_n M)_{i_1 \ldots i_{n-1} j_n i_{n+1} \ldots i_N} = \sum_{i_n} x_{i_1 \ldots i_{n-1} i_n i_{n+1} \ldots i_N} \times m_{j_n i_n}$ Fig. 2.4 offers an example of $(X \times_1 M)$.

We denote the outer-product of two vectors u and v by $u \circ v$.

2.2 Tensor representation of folksonomies

A folkosonomy is generated by a process where users label objects (web-resources uniquely indentified by their urls) with tags. Each user can tag multiple objects, and each object can be tagged by multiple users. The resulting collection of user-object-tag triplets (u_i,o_j,t_k) can be represented by a 3rd-order tensor $\bar{X} \in R^{U \times T \times O}$ or for short , $X_{U \times T \times O}$, where U, T, O respectively denote the sets of users, objects and tags.

The tensor representation of a folksonomy corresponds to a hypergraph linking users, objects, and tags. Similarly to the work in Yeung et al. (2007), we can reduce the resulting hypergraph to a set of three simplified graphs that relate users and objects (UO), users and tags (UT) and tags and objects (TO):

$UO = \langle U \times O, E_{uo} \rangle$, $E_{uo} = \{(u, o) | \exists t \in T : (u, o, t) \in E\}$

$UT = \langle U \times T, E_{ut} \rangle$, $E_{ut} = \{(u, t) | \exists o \in O : (u, o, t) \in E\}$

$TO = \langle T \times O, E_{to} \rangle$, $E_{to} = \{(t,o) | \exists u \in U : (u,o,t) \in E\}$

Previous work on analysis of folksonomies has investigated application of PCA Paolillo and Penumarthy (2007), SVD Prieur et al. (2008), spectral clustering Begelman et al. (2006) to analysis of the resulting matrices. Our proposed methods work directly in the natural representation of folksonomies. In order to leverage this advantage we need first to formally define our learning objectives.

2.3 Ontology Extraction from Folksonomies

Ontologies are formal models of the state of knowledge within a particular domain. The types of knowledge representation languages that are employed range from category systems to taxonomies (Kang et al. (2004)), concept maps (Ausubel et al. (1978)), Galois lattices (Ganter and Wille (1998)) and Descrition Logics (Baader (2003); Bao et al. (2006)). In the current paper we opted for extracting ontologies in the simple form of category systems:

Category systems We define a category as a non-empty set of tags. The set of all categories associated with a set of tags T is denoted by C(T).

$$C(T) = \{c | \exists t \in T, s.t. t \in c\} \tag{2.1}$$

A category system is a set of categories. There are two types of category systems:

- Hard-assignment category systems: A hard-assignment category system corresponding to a given set of tags T is a subset of categories from $C(T)$ that is both *covering* and *non-overlapping*. The *coverage* requirement means that the category system includes all of the original tags and the *non-overlapping* requirement means that each tag appears only in one category:

  $$C_h(T) = \{c_i | \forall t \in T, \exists c_i \in C_h(T) s.t. t \in c_i \wedge \nexists j \neq i s.t. t \in c_j\} \tag{2.2}$$

  The set of all hard-assignment category system is denoted by $C_H(T)$.

- Soft-assignment category systems: In the case of multidimensional datasets the hard-assignment requirement becomes too restrictive: for example, a user is required to be part of a single category. Soft-assignment category systems drop these constraints and become a superset of the hard-assignment category systems. The set of all soft-assignment category systems corresponding to a given set of tags T is denoted by $C_S(T)$ and is the powerset of $C(T)$:

$$C_S(T) = P(C(T)) \tag{2.3}$$

2.4 Community Discovery from Folksonomies

Folkonomies offer a rich source of information for discovering communities of users that share similar interests (as reflected by the objects they tag) and attitudes (as reflected by the specific tags that they use to label the objects. A natural definition of community emerging from a folksonomy is that of a set of individuals (users) that share more of their vocabulary(tags) and preferences(objects) with the community than with the outside world. More formally, $K(\bar{X})$ is a community defined by its categories of users (U_k), tags (T_k) and objects (O_k) from the associated folksonomy $\bar{X}$ if:

$$1 \geq \frac{|E(U_k) \cap E(T_k) \cap E(O_k)|}{|E(U_k) \cup E(T_k) \cup E(O_k)|} > \delta \tag{2.4}$$

Where:

$K(\bar{X})$ - tensor describing the relations inside the cluster

U_k - set of users

T_k - set of tags

O_k - set of objects

$E(.)$ - the set of rows, columns, tubes that contain a particular user,tag,object

$|.|$ - the cardinality of a set

δ - the threshold value (tipically 0.5)

CHAPTER 3. DECOMPOSITION METHODS

In the following we present the matrix decomposition methods which are the building blocks for most of the clustering methods presented. Once presented, we move to their tensor generalizations and to the methods for computing them.

3.1 Two-dimensional decompositions

Matrix decomposition methods express the original matrices as a product of factors in a canonical form. One of the most useful decompositions for data analysis is the Singular Value Decomposition:

The Singular Value and the Rank Decomposition

Definition. SVD re-writes a matrix M as the product of two matrices representing its left and its right orthonormal basis and a diagonal matrix of singular values of M:

$$M = UDV' \tag{3.1}$$

where:

M is a m-by-n matrix with elements from $\mathcal{R}$

U is a m-by-p matrix (left eigenvectors, row space)

D is a p-by-p diagonal matrix (eigenvalues)

V’ is the transpose of the n-by-p matrix V (right eigenvectors, column space)

Eigenvalue Decomposition definition and its relation with SVD An eigenvector x is a vector who’s direction is either unchanged by the transformation (for positive eigenvalues) or reversed (for negative eigenvalues). By denoting the set of all eigenvectors with X, the set of all eigenvalues with the diagonal matrix D and the transformation with M, we obtain:

$$MX = DX \tag{3.2}$$

By simple matrix manipulation we observe that SVD and eigenvalue decomposition are closely related:

$M'M = (VD'U')(UDV') = V(D'D)V'$

$MM' = (UDV')(VD'U') = U(DD')U'$

By replacing V by X in the first equation and U by X in the second equation we find that the right hand sides of these relations describe the eigenvalue decompositions of the left hand sides. Furthermore, the left eigenvectors of M which are contained in U are the eigenvectors of MM' and the right eigenvectors of M which are contained in V are the eigenvectors of $M'M$.

Rank Decomposition definition and its relation with SVD. A matrix rank decomposition writes the matrix as a sum of lower rank matrices. For 2-dimensional matrices, SVD is also a rank-one decomposition:

$$M = \sum_{i=1,n} \lambda_i u_i v_i' \tag{3.3}$$

SVD as a dimensionality reduction method. Theorem. (From: Trefethen and Bau, III (1997)) Let the m-by-n matrix M with the singular values $\lambda_1 > \lambda_2 > ... > \lambda_n$ (some of them can be zero) have rank r. Then, for any $k < r$, it can be shown that the best L_2-norm rank k approximation of M is $K = \sum_{i=1,k} \lambda_i u_i v_i'$:

$$K = argmin_N(k)||M - N||_F \tag{3.4}$$

where $\|.\|_F$ is the Frobenius Norm of the matrix.

Because of the theorem above, SVD is at the core of many clustering and data reduction techniques for domains where the data represents binary relationships between two types of elements (e.g., users and tags). In the case of only one type of instances, the relations might be symmetrical and the special case of eigen-decomposition can be used. For two types of instances, SVD is needed. If there are more than two types of instances which are still related by binary relations, we get the problem of multipartite clustering, which is an active research area (See: Long et al. (2006)). Finally, for unrestricted N-ary relations, we have tensor decomposition methods.

3.2 Higher-order decompositions

Folksonomies represent ternary or 3-way relationships between users, objects, and tags. Consequently, SVD cannot be directly applied to decomposition of folksonomies. One approach to analysis of folksonomies is to first *project* the tensor representing the ternary relationships between users, tags, and objects in a folksonomy into 2-way relationships and then apply standard singular value decomposition (SVD) to the resulting 2-dimensional matrices Wu et al. (2006). In a multi-way setting the reliance on projecting a higher order matrix into 2-dimensional matrices makes the results difficult to interpret. In what follows, we describe PARAFAC (Harshman and Lundy (1994); Carroll and Chang (1970)), Tucker3 (Tucker (1966)) and HOSVD (Vasilescu and Terzopoulos (2002)) which can be viewed as generalizations of SVD to the setting with multi-way relationships.

3.2.1 PARAFAC/CANDECOMP

Definition. Properties Parallel Factor Decomposition (PARAFAC - Harshman and Lundy (1994)) is also known as *Canonical Decomposition* (CANDECOMP - Carroll and Chang (1970)). PARAFAC generalizes SVD to tensors of order greater than 2 (See Figure 3.1)

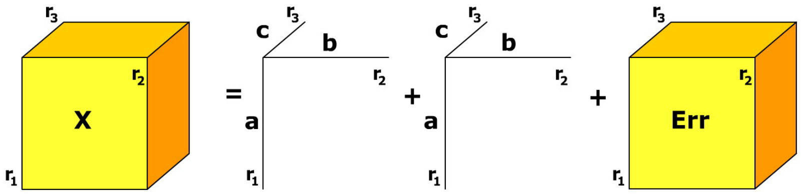

Figure 3.1 The PARAFAC model

In the case of a 3rd-order tensor $\bar{X}$, PARAFAC approximates $\bar{X}$ using the sum of the first w of the outerproducts (*triads*) of vectors and a tensor denoted by $\bar{E}$, representing the *error* of the approximation relative to the tensor $\bar{X}$ (also shown in Algorithm 1):

$$\bar{X} = \sum_{r=1,F} \lambda_r u_r \circ t_r \circ o_r + \bar{E} \tag{3.5}$$

In turn, each element of the cube $\bar{X}$ can be written as:

$$\hat{x}_{ijk} = \sum_{l=1,w} u_{il} t_{jl} o_{kl} \tag{3.6}$$

Algorithm The PARAFAC model is computed by *Alternating Least Squares* (ALS - Leeuw et al. (1976)). For a tensor of order three, ALS estimates the matrices U, T, and O by iterating between the least-squares estimates for one of the matrices while keeping the other two matrices fixed at their most recent estimates.

Algorithm 1 PARAFAC($\bar{X}, w, \epsilon$)

```
Require: 3-dimensional tensor $\bar{X}$, number of components $w$ and minimum improvement step
   $\epsilon$.
Ensure: $U$,$T$,O components.
Initialize T and O
while ($\Delta L > \epsilon$) do
    $U = X_{(1)} Z' (ZZ')^{-1}$ where $Z$=T $\otimes$ $O$ and $X_{(1)}$=unfold($\bar{X}$,1)
    $T = X_{(2)} Z' (ZZ')^{-1}$ where $Z$=U $\otimes$ $O$ and $X_{(2}$=unfold($\bar{X}$,2)
    $O = X_{(3)} Z' (ZZ')^{-1}$ where $Z$=U $\otimes$ $T$ and $X_{(3)}$=unfold($\bar{X}$,3)
    $L(U,T,O) = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{p} \left( x_{ijk} - \sum_{f=1}^{F} u_{if} t_{jf} o_{kf} \right)^2$
end while
```

A limitation of PARAFAC is that it requires the same number of components to be extracted for each of the modes. PARAFAC can be viewed as a constrained version of Tucker3, a technique for 3-way decomposition of a 3rd order tensor. Neither PARAFAC nor Tucker3 require orthogonality of the decomposition. When such orthogonality is desired, HOSVD (Tucker (1966)), a variant of Tucker3 decomposition with orthogonality constraints can be used.

3.2.2 Tucker3

Definition. Properties Tucker3 differs from the PARAFAC model is that it allows the extraction of different numbers of components for each of the modes of the tensor. Figure 2 illustrates the Tucker3 decomposition. The tensor $G_{(W_1 \times W_2 \times W_3)}$ determines the weights assigned to the different components.

$$\hat{x}_{ijk} = \sum_{l=1,w_1} \sum_{m=1,w_2} \sum_{n=1,w_3} a_{il} b_{jm} c_{kn} g_{lmn} \tag{3.7}$$

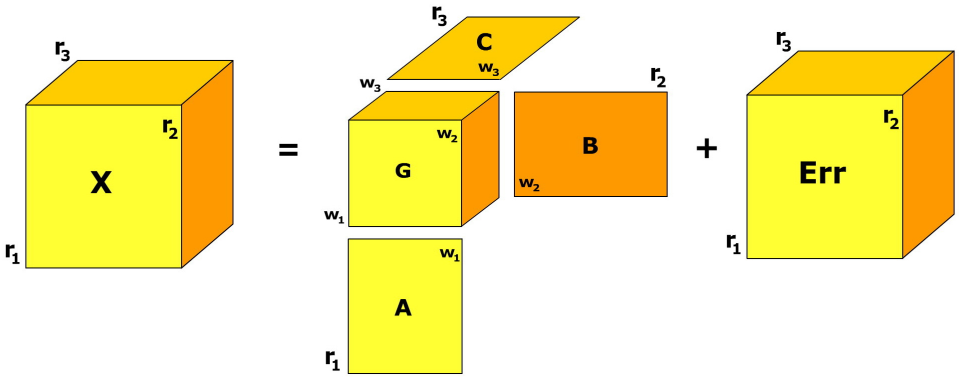

Figure 3.2 The TUCKER3 model

Algorithm. Tucker3 (Tucker (1966)) is computed similarly to PARAFAC by *Alternating Least Squares*:

3.2.3 HOSVD

Definition. Properties. HOSVD or N-mode SVD is the natural extension of SVD to higher-order tensors. Since a matrix is a tensor of order 2, the SVD decomposition of a matrix M ($M = UDV'$) can be also expressed in tensor notation: $M = D \times_1 U \times_2 V$ where U and V are orthogonal 1-mode and 2-mode spaces. By extension, the generalization of SVD for an Nth-order tensor is a *mode-n product* of N orthogonal spaces and the core tensor $\bar{G}$:

$$\bar{X} = \bar{G} \times_1 U_{(1)} \times_2 U_{(2)} \times_3 ... \times_N U_{(N)} \tag{3.8}$$

Algorithm 2 TUCKER($\bar{X}, w_1, w_2, w_3, \epsilon$)

```
Require: 3-dimensional tensor $\bar{X}$, number of components $w_1, w_2, w_3$ and minimum improvement step $\epsilon$.
Ensure: U,T,O,$G = U'X(O \otimes T)$
Initialize T and O using SVD
while ($\Delta L > \epsilon$) do
    U=first $w_1$ left singular eigenvectors of $X_{(I \times JK)}(T \otimes O)$
    T=first $w_2$ left singular eigenvectors of $X_{(J \times IK)}(O \otimes U)$
    O=first $w_3$ left singular eigenvectors of $X_{(K \times IJ)}(U \otimes T)$
    $L(U,T,O) = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{p}(x_{ijk} - \hat{x}_{ijk})^2$
end while
```

where each $U_{(i)}$ contains the eigenvectors of the column space of the matrix $X_{(i)}$ which is obtained by unfolding the tensor $\bar{X}$ on the mode i.

HOSVD in matrix notation

$$X_{(i)} = U_{(i)}Z_{(i)}(U_{(1)} \otimes U_{(2)} \otimes ... \otimes U_{(N)}) \tag{3.9}$$

Expressing the HO-SVD on each one of the unfoldings $X_{(i)} = U_{(i)} \times G_{(i)}(U_{(1)} \otimes ... \otimes U_{(N)})$ can be viewed as standard PCA where the matrix $U_{(i)}$ is the matrix of eigenvectors and loadings matrix is obtained as the product of the unfolding of the core tensor $G_{(i)}$ times the Kronecker product of the other unfoldings.

Ways of computing $U_{(i)}$:

$$U_{(i)} = SVD(X_{(i)}X_{(i)}^t) \tag{3.10}$$

$$U_{(i)} = X_{(i)} \times (Z_{(i)}(U_{(1)} \otimes U_{(2)} \otimes ... \otimes U_{(N)}))^{-1} \tag{3.11}$$

The HOSVD algorithm Vasilescu and Terzopoulos (2002):

Algorithm 3 HOSVD($\bar{X}, w_1, w_2, w_3$)

```
Require: 3-dimensional tensor $\bar{X}$, number of components $w_1, w_2, w_3$.
Ensure: $U, T, O, \bar{G}$
U=first p left singular eigenvectors of $X_{(I \times JK)}$
T=first q left singular eigenvectors of $X_{(J \times IK)}$
O=first r left singular eigenvectors of $X_{(K \times IJ)}$
$\bar{G} = \bar{X} \times_1 U \times_2 T \times_3 O$
```

3.3 Comparative model complexities for two and three-dimensional decompositions

The SVD decomposition with k components of a $U \times WT$ matrix obtained by unfolding the three-dimensional tensor $\bar{X}^{U \times T \times O}$ requires $k(U + TO)$ parameters. By comparison, a corresponding Tucker3 model will need $k(U + T + O) + k^3$ parameters, and the PARAFAC model only $k(U + T + O)$. PARAFAC can be viewed as a constrained version of Tucker3, and Tucker3 a constrained version of two-way PCA Kiers and Van Mechelen (2001). Hence, any data set that can be modeled adequately with PARAFAC can thus also be modeled by Tucker3 or 2-way PCA of unfolded matrices. However, PARAFAC uses fewer degrees of freedom to model the data, and hence is attractive from the standpoint of Occam's razor.

CHAPTER 4. CLUSTERING USING DECOMPOSITION METHODS

We now proceed to describe several methods for clustering the user-tag-object triplets of a folksonomy tensor in order to extract ontologies and user communities.

4.1 2D Clustering and Analysis Methods

In the following we review the state-of-the art methods in clustering and dimensionality reduction. In the next subsection we will draw equivalence relations between them and discuss their implications from the point of view of clustering.

4.1.1 K-means.

K-means (In Hartigan and Wong (1979)) is one of the most well-known clustering methods due to its conceptual and implementation simplicity, its scalability and performance. K-means method assumes as known the number of clusters (k) and computes their centroids in an iterative fashion by minimizing the sum of squared distances between the centroids and the datapoints in the respective clusters:

$$J_K = \sum_{k=1,K} \sum_{i \in C_k} (x_i - m_k)^2 \tag{4.1}$$

where:

$X = (x_1, ...x_n)$ is the matrix containing the datapoints

$m_k = \sum_{i \in C_k} \frac{x_i}{n_k}$ is the centroid of cluster k

n_k is the number of datapoints in cluster k

K-means is also very sensitive to the choice of starting conditions, since due to its iterative fashion it is prone to get stuck in local minima.

4.1.2 Principal Component Analysis (PCA)

Principal Component Analysis (Pearson (1901)) is a very popular dimensionality reduction method. It assumes that the data lies in an Euclidean space and that implicitly the dependencies between attributes are described only in terms of the first order-statistics, namely their covariance matrix. PCA finds an orthonormal basis of the initial data either by eigendecomposition of the covariance data, either by singular value decomposition of the mean normalized dataset:

Let X be the original data matrix and $Y = (y_1, y_2, ...y_n)$, $y_i = x_i - \bar{x}$, $\bar{x} = \sum_i x_i/n$ be the centered data matrix.

Let $UDV' = Y$ be the SVD of Y. Then Y can be written also as: $Y = \sum_k \lambda_k^{1/2} u_k v'_k$. Using SVD-approximation theorem presented in Section 3, it can be easily shown that, in the transformed space, the principal eigenvector contains the maximum variance possible for a projection, the second eigenvector contains the second greatest variance, and so on. The eigenvectors with low variance can be discarded and optimal dimensionality reduction (in terms of L_2-norm) is achieved.

4.1.3 Spectral clustering.

In spectral clustering (Shi and Malik (2000)), the data is represented in the form of a weighted graph where the nodes represent data points and edge weights represent their pairwise similarities. The graph is characterized by its adjacency matrix A. If the goal of clustering is to minimize the similarity among data points assigned to different clusters then this is equivalent in graph-theoretic terms to finding the minimum cut of the graph. If we denote the disjoint sets of nodes of the two resulting subgraphs by A and B and $w(u, v)$ is the weight of the edge connecting u and v, we have the following optimization:

$$min_{A,B} cut(A, B) = \sum_{u \in A, v \in B} w(u, v) \tag{4.2}$$

The disadvantage of minimum graph cut as a clustering objective is that it favors the erosion of the graph, leading to the discovery of small clusters. In order to encourage the formation of more balanced clusters a new clustering objective can be derived, namely *Normalized Cut*:

$$Ncut(A,B) = \frac{cut(A,B)}{vol(A)} + \frac{cut(A,B)}{vol(B)} \tag{4.3}$$

where:

$vol(X) = \sum_{x \in X, y \in V} w(x,y)$ Another equivalent clustering objective is:

$$Nassoc(A,B) = \frac{assoc(A,A)}{vol(A)} + \frac{assoc(B,B)}{vol(B)} \tag{4.4}$$

where:

$assoc(X,Y) = \sum x \in X, y \in Y w(x,y)$

$$Ncut(A,B) = 2 - Nassoc(A,B) \tag{4.5}$$

By matrix manipulations, (see Shi and Malik (2000)), it can be shown that minimizing the Ncut objective is equivalent to computing the minimum of the *Rayleigh quotient* (Golub and Van Loan (1996)):

$$min_x Ncut(x) = min_y \frac{y^T D^{-1/2}(D-W)D^{-1/2}y}{y^T y} \tag{4.6}$$

where:

D is the diagonal outdegree matrix: $d_{(1)ii} = \sum_{k=1,N} a_{ki}$:

$$L = D^{-1/2}(D-W)D^{-1/2} \tag{4.7}$$

is called the *Normalized Laplacian* of graph G. From literature (Golub and Van Loan (1996)), we know that the *Rayleigh quotient* si minimized by the smallest eigenvector of L. However, the smallest eigenvector of L is always 1, so the non-trivial solution to the optimization problem is the second smallest eigenvector.

In the case bipartite graphs, the Laplacian can be generalized like in Dhillon (2001) to:

$$L = D_{(1)}^{-1/2} A D_{(2)}^{-1/2}, \tag{4.8}$$

where: $D_{(1)}$ is the diagonal outdegree matrix and $D_{(2)}$ is the diagonal indegree matrix: $d_{(2)ii} = \sum_{k=1,N} a_{ik}$.

4.2 Equivalencies and Clustering strategies.

4.2.1 The equivalence between K-means and PCA

From: Ding and He (2004). The authors show that the principal components of an initial data matrix X are also the relaxed solution of the cluster membership indicators in K-means clustering. The traditional and the newly discovered derivation of PCA are not contradictory: clustering is in fact a form of data reduction to the cluster space by comparison with SVD where the mapping is to an Euclidean space. The paper shows that in fact the space spanned by the principal directions according to PCA is identical to the space specified by the between-clusters scatter matrix.

Proof: K-means tries to minimize the following eq.:

$$J_k = \sum_{k=1,K} \sum_{i \in C_k} (x_i - m_k)^2 \tag{4.9}$$

We define the distance between two clusters in terms of their components' Euclidian distances:

$$d(C_k, C_l) = \sum_{i \in C_k} \sum_{j \in C_l} (x_i - x_j)^2 \tag{4.10}$$

We can rewrite J_k as a function of clusters distances:

$$J_k = \sum_{k=1,K} \sum_{i,j \in C_k} (x_i - x_j)/2n_k = n\bar{y^2} - 1/2J_D \tag{4.11}$$

Where $\bar{y^2} = \sum_i y_i' y_i / n$ and J_D is equal to:

$$J_D = n_1 n_2/n[2d(C_1, C_2)/n_1 n_2 - d(C_1, C_1)/{n_1}^2 - d(C_2, C_2)/{n_2}^2] \tag{4.12}$$

Therefore: $min(J_k) \equiv max(J_D)$. Furthermore:

$$J_D = -q' D q \tag{4.13}$$

Where q is the indicator vector:

$$q(i) = \sqrt{n_2/nn_1}, ifi \in C_1, -\sqrt{n_1/nn_2}, ifi \in C_2 \tag{4.14}$$

and D is the pairwise distances matrix:

$$d_{ij} = \|x_i - x_j\|^2 \tag{4.15}$$

A classical result from matrix algebra is that the continuous solution of q for minimizing $J(q) = q'Dq/q'q$ (the Rayleigh quotient) is the eigenvector corresponding to the smallest non-zero eigenvalue of D:

$$\lambda^* = argmin_q q'Dq/q'q \tag{4.16}$$

The proof is complete.

4.2.2 The equivalence between Kernel K-means, Kernel PCA and Spectral Clustering

. Similarly to K-means and PCA, Kernel K-means and Kernel PCA can be proved to be equivalent. Another interesting result from Dhillon et al. (2004) extends the equivalence between Kernel K-means and Kernel PCA to spectral clustering. The authors introduce a new weighted form of Kernel K-means that minimizes the following equation (where $\phi(.)$ is the nonlinear kernel function):

$$J_k^{kernel} = \sum_{k=1,K} \sum_{i \in C_k} w(i) \|\phi(i) - m_k\|^2 \tag{4.17}$$

Where: m_k is a weighted centroid of cluster k: $m_k = \frac{\sum_{i \in C_k} w(i)\phi(i)}{\sum_{i \in C_k} w(i)}$ Let π_k be a cluster indicator and $d(\pi_k) = \sum_{i \in C_k} w(i)\|\phi(i) - m_k\|^2$ be the *distortion* of cluster k according to π_k. Let $s_k = \sum_{i \in C_k} w(i)$ be the sum of the weights of points in a cluster. Let W be the weight matrix for all clusters and W_k the diagonal matrix corresponding to weights used in cluster k. We then have that:

$$m_k = \phi_k \frac{W_k e}{s_k} \tag{4.18}$$

Where: e is the unity vector.

The distortion vector becomes: $d(\pi_k) = \sum_{i \in C_k} w(i)\|\phi(i) - \phi_k \frac{W_k e}{s_k}\|^2 =$
$= \|(\phi_k W_k^{1/2}(I - \frac{W_k^{1/2} ee' W_k^{1/2}}{s_k})$.

We observe that $P = (I - \frac{W_k^{1/2} ee' W_k^{1/2}}{s_k})$ is an orthogonal projection, i.e.$P = P^2$. We then have that:

$d(\pi_k) = trace(W_k^{1/2}\phi_k^T\phi_k W_k^{1/2}) - \frac{e^T W_k}{\sqrt{s_k}}\phi_k^T\phi_k\frac{W_k e}{\sqrt{s_k}}$

For the J_k^{kernel} we have that: $J_k^{kernel} = trace(W^{1/2}\phi^T\phi W^{1/2})$ - $trace(Y^T W^{1/2}\phi^T\phi W^{1/2} Y)$ where Y is a diagonal matrix with elements: $\frac{W_i e}{\sqrt{s_i}}$ This is equivalent to the following maximization:

$$max_Y trace(Y' W^{1/2} K W^{1/2} Y) \tag{4.19}$$

Where: K is the kernel matrix $K = \phi'\phi$

It is easy to see this is similar to the Rayleigh quotient, the same formula that determines the *Normalized Cut*. The proof is complete.

4.2.3 The equivalence between Spectral Clustering and Local Dimensionality Reduction

(Belkin and Niyogi (2003)) Let G be the graph associated to the relational data represented by the matrix W. Let $x = (x_1, x_2, x_3...x_n)'$ be a mapping of the original graph G to a reduced representation. A good reduced representation of G is one that preserves its local structure. A corresponding minimization objective can be of the form:

$$\sum_{ij}(x_i - x_j)^2 W_{ij} \tag{4.20}$$

Minimizing the above objective would insure that originally close a_i and a_j will have close mappings represented by x_i and x_j. We can re-write the above formula as:

$$\frac{1}{2}\sum_{ij}(x_i - x_j)W_{ij} = x^T L x \tag{4.21}$$

The minimization objective becomes:

$$argmin_{x, x^T Dx = 1, x^T D1 = 0} x^T L x \tag{4.22}$$

The restriction $x^T Dx = 1$ removes an arbitrary scaling factor from the embedding. The restriction $x^T D1 = 0$ removes the trivial solution of mapping to 1 (remove the translation invariance in x). Let A and B be disjoint sets of V, $A \cup B = V$ and $a = vol(A)$ and $b =$

$vol(b)$.Let: $x_i = \frac{1}{a}$ if $x_i \in A$, $\frac{1}{b}$ if $x_i \in B$.We have that:

$$x^T Lx = \sum_{ij}(x_i - x_j)^2 W_{ij} = \sum_{V_i \in A, V_j \in B} (\frac{1}{a} + \frac{1}{b})^2 cut(A, B) \tag{4.23}$$

Also:

$$x^T Dx = \sum_i x_{ii}^2 d_{ii} = \sum_{V_i \in A} \frac{1}{a^2} d_{ii} + \sum V_i \in B \frac{1}{b^2} d_{ii} = \frac{1}{a^2} vol(A) + \frac{1}{b^2} vol(B) = \frac{1}{a} + \frac{1}{b} \tag{4.24}$$

Therefore:

$$\frac{x^T Lx}{x^T Dx} = cut(A, B)(\frac{1}{a} + \frac{1}{b}) = Ncut(A, B) \tag{4.25}$$

Let $y = D^{1/2} x$:

$$\frac{x^T Lx}{x^T Dx} = \frac{y^T D^{1/2} L D^{1/2} y}{y^T y} \tag{4.26}$$

Let $\hat{L} = D^{1/2} L D^{1/2}$ be the Normalized Laplacian.

Note:$Lx = \lambda Dx$, $D^{1/2}(D - W)^{1/2} z = \lambda z$, $D^{-1} W y = \lambda y$ all have the same eigenvectors. Thus Spectral Clustering and Local Embedding are equivalent.

4.2.4 Clustering Strategies

In the previous section we saw that depending on the matrix to be decomposed there are different ways in which the principal eigenvectors can be used for clustering. These strategies are for short:

- The eigenvectors of the original data matrix can be used directly as cluster indicators with cut-off at zero resulting in soft clustering.

- Because of the equivalence of Spectral Clustering with Kernel PCA, the eigenvectors of the Normalized Laplacian can also be used directly as cluster indicators with cut-off at zero resulting in soft clustering.

- The eigenvectors of the Laplacian matrix can be used indirectly for clustering in conjunction with the standard K-means algorithm.

4.3 Higher-order Clustering

In the case of folksonomies, the dataset is a tensor (in graph-theoretic terms, it is a hypergraph). Consequently, we need to extend spectral clustering techniques to work with tensors instead of matrices.

4.3.1 Clustering by decomposition of adjacency tensor

PARAFAC: Similarly to the SVD of the two-dimensional data being interpreted as the solution of the relaxed K-means, higher-order tensor decompositions can be used directly on the data tensor to discover clusters: A PARAFAC model of rank R decomposes the original data tensor in R clusters, each having its relative importance given by λ_r. Formally we write the resulting clusters Y_r as following: $Y_r^{Parafac} = (u_r, t_r, o_r), r = 1, R$. Therefore each group has three facets: its users(community), its tags(vocabulary) and objects.

TUCKER3 and HOSVD: Even if the PARAFAC model has the advantage of being relatively fast to compute, it has a series of short-comings: it is sometimes numerically unstable, the principal components are not generally orthonormal and the number of components are the same across all the three dimensions. In the case of TUCKER3 and HOSVD decompositions, the groupings are obtained by decreasingly ordering the core values and their corresponding triplets of principal components. The resulting groupings $Y_r^{Tucker3} = (u_r, t_r, o_r), r = 1, R$ have a similar interpretation to the ones obtained by PARAFAC.

K-means: The input for the K-means algorithm for each of the points from the original three-dimensional space is the union of the three k-approximations of the original dimensions. Each of the points is hard-assigned to one of the clusters.

4.3.2 Clustering by decomposition of Laplacian tensor

For a bidimensional tensor with different row and column spaces the Normalized Laplacian is defined as $L = D_1^{-1/2} A D_2^{-1/2}$, which can be written in tensor notation as:

$$L = A \times_{(1)} D_1^{-1/2} \times_{(2)} D_2^{-1/2} \tag{4.27}$$

For a three-dimensional tensor the straight-forward generalization is:

$$L = X \times_{(1)} D_1^{-1/3} \times_{(2)} D_2^{-1/3} \times_{(3)} D_3^{-1/3} \tag{4.28}$$

Using the HOSVD algorithm to decompose the normalized tensor is equivalent to computing U,T,O as:

$$U = SVD(D_1^{-1/3} X_{(I \times JK)} D_{23}^{-1/3}, p) \tag{4.29}$$

$$T = SVD(D_2^{-1/3} X_{(J \times IK)} D_{13}^{-1/3}, q) \tag{4.30}$$

$$O = SVD(D_3^{-1/3} X_{(K \times IJ)} D_{12}^{-1/3}, r) \tag{4.31}$$

where: SVD(X,i) returns the first i left singular eigenvectors of X.

Note: These equations are very similar to the Normalized Laplacian for bipartite graphs, but with the different normalization factors.

In order to handle an arbitrary number of clusters, the cluster indicators can be obtained by case either directly from the first principal components, either recursively by dividing the first initial two clusters in finer-grained clusters, either using K-means as explained above.

CHAPTER 5. RESULTS

5.1 Datasets

We compare the performance of two-dimensional methods with multi-dimensional methods on two real-word datasets, both produced by collective/social bookmarking websites:

- The data in the first dataset was gathered by crawling the social bookmarking website del.ici.o.us. The final dataset consists of 7776 relating 588 users, 203 tags and 693 websites. The dataset was obtained by crawling the website in a tag-first manner, collecting the first 100 pages and the associated users for each tag, by breadth-first search. The starting tag was chosen to be ”Web2.0”.

- The data in the second dataset is a snapshot of the entire bookmarking history of the research-oriented website citeulike.org. The website allows researchers to bookmark papers of interest with tags and make them public to the community. The snapshot covers the period 2004-11 to 2008-02 and it is partly filtered for junk links and tags. It contains 727,542 unique papers, 22,562 users and 152,296 tags combined in 2,398,267 triplets.

5.2 Preprocessing. Graph Erosion

In both of the real-world datasets we observed a power-law distribution of the objects, tags and users. The idea behind the preprocessing step was to remove part of the tail that contains noisy signal and to keep objects, users and tags that share reliable patterns of connectivity. We call the following pre-processing method *graph erosion*: For each dimension we choose a minimum number of triplets that it needs to appear in order to be kept in the dataset. In one path through the dataset we remove the triplets that contain objects, users or tags that have counts less than the required threshold. We update the counts for each user, tag and object based on the remaining triplets and repeat the process until no more triplets are removed. It is easy to see that loosely connected components are the ones that disappear and that is why the method is similar to an erosion. The components that are guaranteed to be maintained are the ones that share connections within themselves more than the threshold values for each dimension. In our experiments we employ a threshold of at least 3 triplets for all of the three dimensions in the case of the Del.icio.us dataset and a threshold of 10 triplets in the case of the Citeulike dataset. This reduces the Del.icio.us dataset to 7372 triplets connecting 574 websites, 450 tags and 175 users and the Citeulike dataset to 454,230 triplets connecting 21935 articles, 5250 tags and 1550 users.

5.3 Measuring performance

5.3.1 Cluster purity measure

We chose to use a clustering score that is inverse proportional with the normalized cut Shi and Malik (2000), which we call *Weighted Purity* (P) and measures for each cluster the proportion of the selected items in-links in their total number of links:

Computing the average P-score for 3-dimensional clustering:

$$P = \frac{\sum_{i=1,k} P_i}{k} \tag{5.1}$$

where P_i is the weighted purity for the i-th cluster:

$$P_i = \frac{|E(tags_i) \cap E(users_i) \cap E(objects_i)|}{|E(tags_i) \cup E(users_i) \cup \{E(objects_i)|} \tag{5.2}$$

Where:

$\{E(x)\}$ is the set of all hyperedges that have one of their extremities in x.

In order to make the P-measure of the simplified model compatible with the three-dimensional P-measure, for each of the three simplified views (TO,TU,OU) we infer for each cluster the elements of the missing dimension by retrieving all of the triplets containing elements from the known dimensions and returning the set of unique values across the third dimension. Intuitively, this corresponds to labeling each triplet by looking only at two of its dimensions and looking at the values of the third dimension after the fact. For example, assuming that we marginalize the objects dimension we obtain the following P-score:

$$\hat{P}_i = \frac{|E(tags_i) \cap E(users_i)|}{|E(tags_i) \cup E(users_i) \cup \hat{E}(objects_i)|} \tag{5.3}$$

where the set of values in the third dimension is computed:

$$objects_i = value(\{E(tags_i) \cap E(users_i)\}, Dim_{objects}) \tag{5.4}$$

where $value(S_i, dimension_j)$ is the set of all values that the set of triplets S_i takes along the $dimension_j$

5.3.2 Soft clustering efficiency

With the increase of the number of dimensions along which the clusters are described, the requirement of clustering by hard assignment becomes less and less feasible. For example, in the case of folksonomies, hard clustering excludes the possibility that the same users can be part of two communities with different interests and vocabularies. However, when using soft assignment the purity measure is not sufficient to characterize the quality of clustering. By soft assignment, some datapoints might not fall in any of the clusters and other datapoints can be assigned to multiple clusters. In the following, we argue that a soft clustering is performing well if it covers as many of the points in the original dataset with as least redundancy as possible. In the limit, the best performing clustering would be a hard assignment clustering. However, by not restricting the method, we can discover more natural clusters and then compare different segmentations of the initial set of datapoints represented by $\bar{X}$. In the following we define the two components of the soft clustering efficiency measure (E) which are the redundancy and coverage ratios:

Average Redundancy Ratio (R):

$$R(C) = \sum_{x \in D} \frac{|C_x|}{|C|} \tag{5.5}$$

where:

D is the dataset

C is the set of clusters

C_x is the set of clusters that contain the datapoint x.

Coverage Ratio (COV):

$$COV(C) = \frac{|D_C|}{|D|} \tag{5.6}$$

where:

D is the dataset

C is the set of clusters

D_C is the set of of points contained in the clusters

Efficiency (E):

$$E(C) = \frac{R(C)}{COV(C)} \tag{5.7}$$

Maximizing E is equivalent with minimizing the average redundancy ratio R and maximizing the clusters' coverage measure COV.

5.4 Experimental results.

5.4.1 Del.icio.us dataset

Quantitative results: In the present subsection we will use the purity and the efficiency measures for a quantitative comparison of clustering using multi-dimensional methods PARAFAC, TUCKER3 and HOSVD and two-dimensional methods on the Del.icio.us dataset. The results are presented in tables 5.1-5.4, with each cell representing the pair of purity and efficiency measures for a specific number of clusters(2,5,10) on a specific type of matrix (A=adjacency/L=laplacian) using a specific clustering method.

Table 5.1 Del.icio.us: Purity/Efficiency of 3D clustering using direct assignment.

clusters	PARAFAC	TUCKER3	HOSVD
2A	0.6/0.6	0.5/0.66	0.01/0.38
5A	0.48/0.28	0.47/0.26	0.11/0.46
10A	0.3/0.18	0.4/0.14	0.05/0.25
2L	0.6/0.6	0.5/0.66	0.01/0.38
5L	0.48/0.28	0.47/0.26	0.11/0.46
10L	0.3/0.18	0.4/0.14	0.05/0.25

Table 5.2 Del.icio.us: Purity/Efficiency of 2D clustering using direct assignment.

clusters	OT	UO	UT
2A	0.84/0.97	0.84/0.97	0.38/0.02
5A	0.61/0.31	0.52/0.41	0.61/0.33
10A	0.3/0.18	0.4/0.16	0.05/0.14
2L	0.15/0.11	0.46/0.93	0.01/0.38
5L	0.52/0.33	0.73/0.24	0.73/0.24
10L	0.39/0.19	0.49/0.15	0.49/0.16

Cluster visualization: The following images show the first three clusters resulted by direct assignment from the first principal components of three and two-dimensional decompositions:

We can observe that by decomposing only using two of the three components, we cannot find any plausible direction in the original three-dimensional dataset. A second observation is the fact that in 5.6, as expected (Shi and Malik (2000)), the first eigenvector of the matrix Laplacian is not informative. Finally, we can see that the three-dimensional decomposition finds shapes in the dataset, and each component associates a cluster to one extremity. The separation is not pure, but it is potentially better than that obtained by two-dimensional methods. The visualization also confirms that the clustering should be using soft assignment since for example in fig.5.3 the same users belong to two clearly defined clusters that differ in terms of tags and objects.

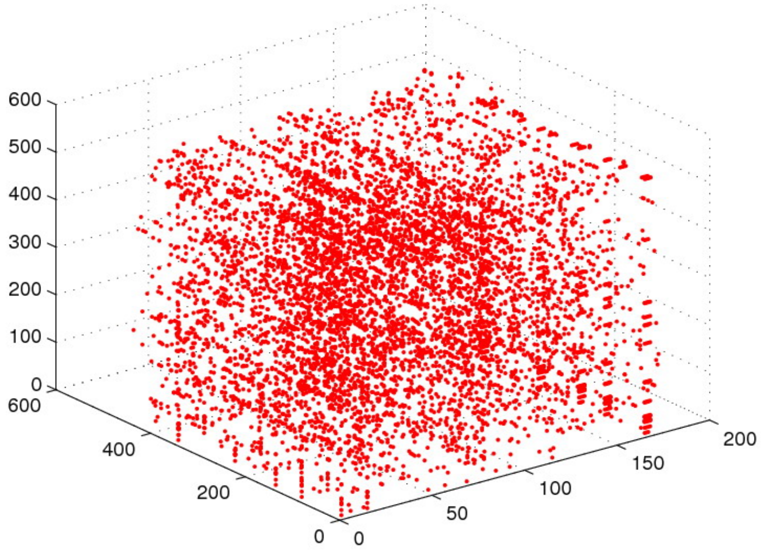

Figure 5.1 Del.icio.us: The original dataset

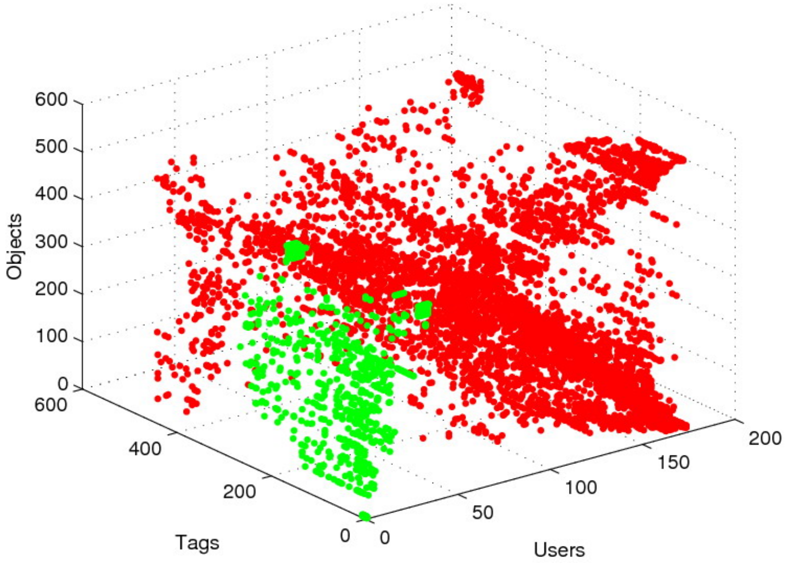

Figure 5.2 3D-Parafac: Cluster no.1

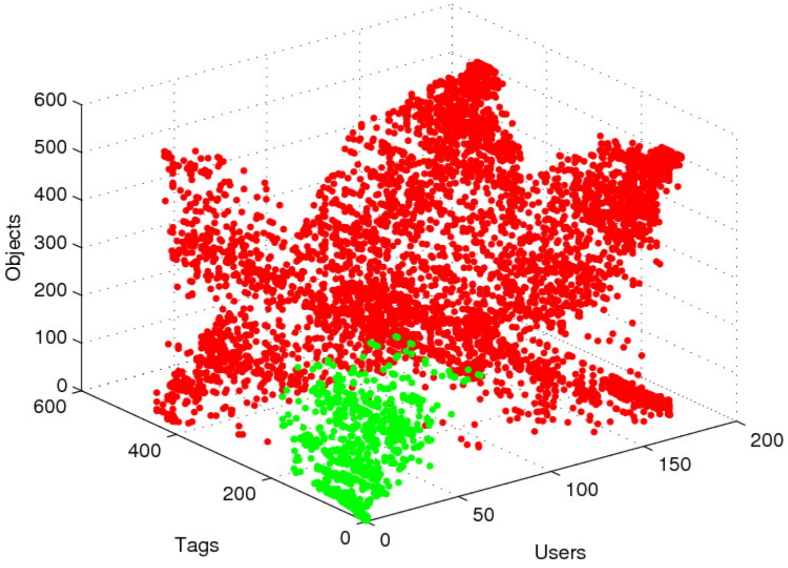

Figure 5.3 3D-Parafac: Cluster no.2

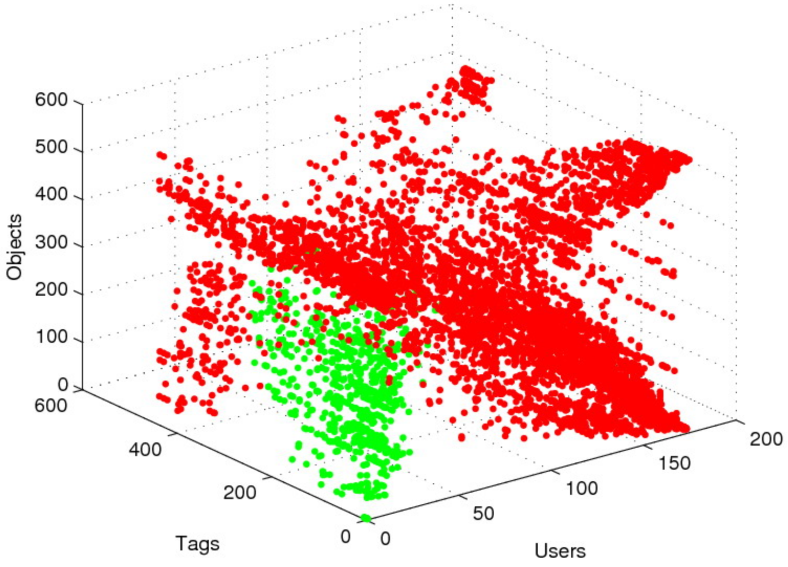

Figure 5.4 3D-Parafac: Cluster no.3

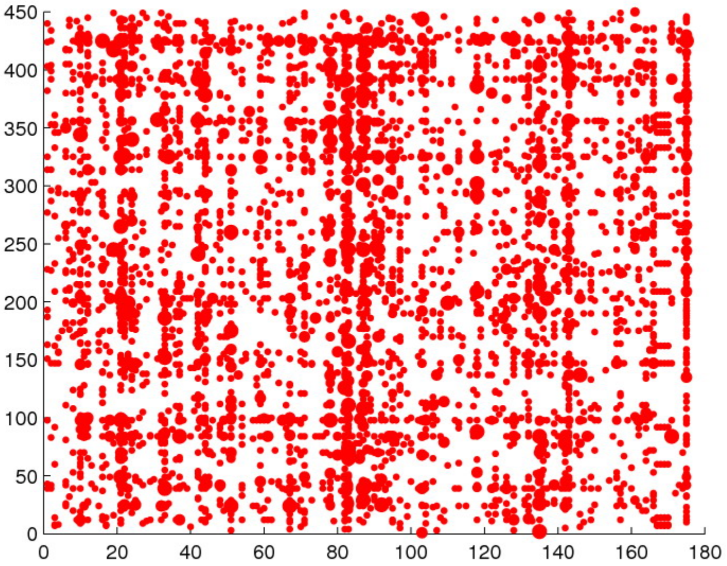

Figure 5.5 The projection of the Delicious dataset on the Object dimension

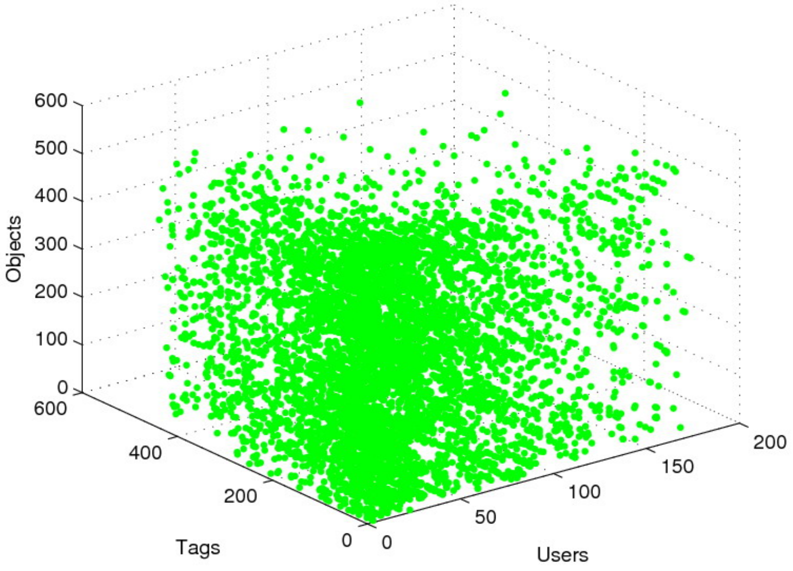

Figure 5.6 2D-SVD of UT: 1st Principal Component

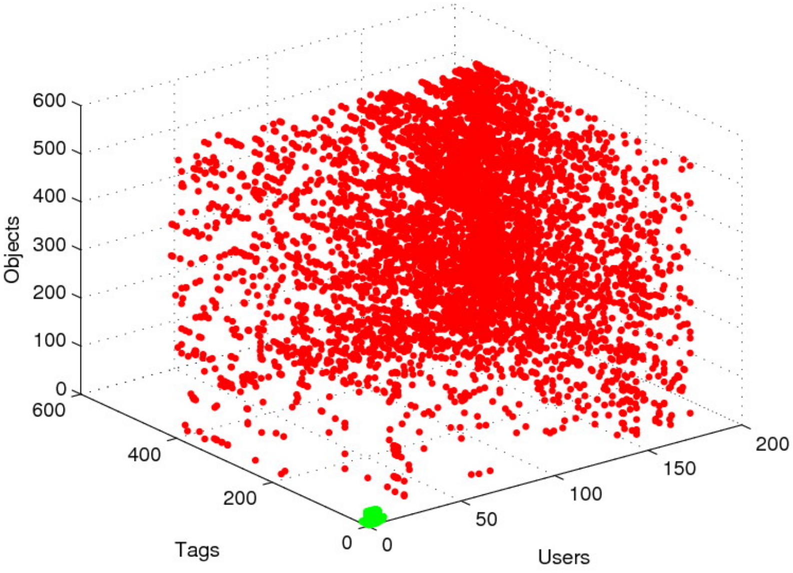

Figure 5.7 2D-SVD of UT: Cluster no.1

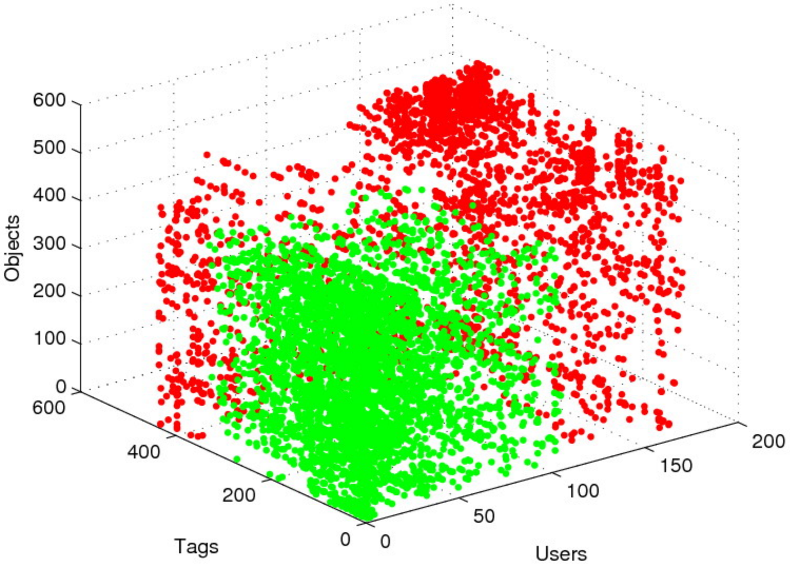

Figure 5.8 2D-SVD of UT: Cluster no.2

Table 5.3 Del.icio.us: Purity/Efficiency of 3D clustering using K-means.

clusters	PARAFAC	TUCKER3
2A	0.85/0.98	0.5/0.98
5A	0.4/0.91	0.42/0.92
10A	0.23/0.83	0.260.85
2L	0.85/0.98	0.85/0.98
5L	0.5/0.92	0.41/0.9
10L	0.24/0.88	0.3/0.84

Table 5.4 Del.icio.us: Purity/Efficiency of 2D clustering using K-means.

clusters	OT	UO	UT
2A	0.56/0.84	0.85/0.98	0.56/0.84
5A	0.4/0.67	0.41/0.68	0.29/0.66
10A	0.29/0.56	0.27/0.56	0.26/0.58
2L	0.72/0.66	0.67/0.68	0.85/0.98
5L	0.53/0.78	0.5/0.65	0.39/0.57
10L	0.31/0.53	0.25/0.45	0.22/0.52

Qualitative results: The qualitative results show the best ranking tags, users and objects in each of the ten clusters. The cluster names are assigned based on the general observed content. In case of ambiguous content the cluster name is followed by ”?”. In case of seemingly random content, the cluster name is denoted by a ”?”. For detailed information on the clusters, see the appendixes 1 and 2.

5.4.2 Citeulike dataset

In the following we compare the quantitative and qualitative results we obtained on the Citeulike dataset using three and two-dimensional clustering methods. The quantitative results are shown in tables 5.6-5.9 and are followed by the qualitative results on the best performing clustering methods in each class.

Qualitative results: In the available dataset, the user ids are anonymized because of privacy concerns. For this reason, we chose not to display the users ids in the qualitative results. For detailed information on the clusters, see the appendixes 3 and 4.

Table 5.5 Del.icio.us: Qualitative results

Cluster no.	PARAFAC	SVD-UT
01	Music	?
02	Auto	Web Design
03	Programming Languages	Web Design
04	Photo Blogging	Media
05	Geo/Travel	Open Source
06	Fashionable	Shopping
07	Directory	Music
08	HowTo/Community	Phone
09	Innovation	Modern Living
10	Web2.0	Web Design

Table 5.6 Citeulike: Purity/Efficiency of 3D clustering using direct assignment.

clusters	PARAFAC	TUCKER3
2A	0.61/0.62	0.32/0.2
5A	0.62/0.28	0.2/0.3
10A	0.34/0.14	0.11/0.22
2L	0.9/0.53	0.4/0.31
5L	0.6/0.25	0.33/0.34
10L	0.47/0.17	0.2/0.14

5.5 Discussion of the results:

5.5.1 Multi versus two-dimensional methods

By comparing the quantitative and qualitative results of three-dimensional versus two-dimensional clustering methods on the folksonomic datasets we draw a series of conclusions: When using the principal components as cluster indicators directly, the two-dimensional methods out-perform multi-dimensional methods from the point of view of quantitative measures. Our hypothesis is that zero is not the right cut-off point for three-dimensional clustering. Quantitatively, the multi-dimensional methods are performing better than two-dimensional methods when using K-means, especially in terms of efficiency. This reinforces the previous explanation of our findings: the three-dimensional decomposition is useful, but mapping it to clustering indicators with cut-off at zero is not optimal. Simple methods such as linear search

Table 5.7 Citeulike: Purity/Efficiency of 2D clustering using direct assignment.

clusters	OT	UO	UT
2A	0.09/0.49	0.29/0.36	0.42/0.78
5A	0.31/0.46	0.42/0.27	0.48/0.43
10A	0.42/0.16	0.38/0.17	0.48/0.23
2L	0.04/0.42	0.28/0.27	0.21/0.18
5L	0.59/0.32	0.63/0.29	0.63/0.36
10L	0.33/0.21	0.58/0.14	0.69/0.17

Table 5.8 Citeulike: Purity/Efficiency of 3D clustering using K-means.

clusters	PARAFAC	TUCKER3
2A	0.68/0.52	0.23/0.24
5A	0.59/0.35	0.31/0.23
10A	0.44/0.24	0.11/0.2
2L	0.89/0.63	0.5/0.41
5L	0.7/0.35	0.41/0.24
10L	0.67/0.14	0.31/0.18

for the best cut-off point in each of the three dimensions can be used to improve it.

We argue that, qualitatively, three-dimensional methods and especially those using the Tensor Laplacian offer much cleaner and exact clusters than two-dimensional methods. This is especially obvious in the Citeulike dataset, where the clusters retrieved using the best of the two-dimensional methods retrieves repeatedly the same papers as being the top in each of the clusters. Some other observations are that not all the dimensions are born equal: Some of the projections prove to be more informative than others, especially the combination User-Tag. The interpretation of the phenomenon is straight-forward: Knowing the users and their vocabulary explains most of their interests (represented by the Object dimension).

Another interesting observation is that the clusters that were retrieved correctly by the two-dimensional methods are different than the ones discovered by the three-dimensional methods. This suggests that the methods might be complementary and can be used together in order to find clusters in folksonomies.

By comparison with the other three-dimensional methods, HOSVD surprisingly under-performed.

Table 5.9 Citeulike: Purity/Efficiency of 2D clustering using K-means.

clusters	OT	UO	UT
2A	0.11/0.31	0.3/0.35	0.39/0.69
5A	0.34/0.42	0.43/0.28	0.5/0.44
10A	0.43/0.18	0.4/0.17	0.49/0.27
2L	0.09/0.42	0.3/0.26	0.23/0.2
5L	0.6/0.29	0.65/0.28	0.65/0.39
10L	0.4/0.19	0.59/0.13	0.68/0.21

Table 5.10 Citeulike: Qualitative results

Cluster no.	PARAFAC	SVD-UT
01	Complex Adaptive Systems	Genetics?
02	Computer Security	Genetics?
03	Education/Pathology	Genetics?
04	Chemistry	Computer/Network Security?
05	Primates	?
06	Alt.Medicine	Electronics
07	Climate change	Administration
08	Mice experiments	Folksonomies
09	Neuropsychology	Folksonomies
10	Orthopedic	Complex Adaptive Systems

For this reason we did not include its results in the subsequent comparisons.

PARAFAC outperforms all other multi-dimensional methods in both in terms of results and scalability and speed. Tucker3 gives results close to those of PARAFAC but in its current implementation (Bader and Kolda (2007)) it requires a lot more computational resources.

5.5.2 Relevance of the results for the task of ontology extraction

Due to the relative purity of the top elements obtained in our clusters, we are optimistic in using three-dimensional clustering as a tool for extracting tag-ontologies in the shape of meaningful category systems. For example, as seen in the qualitative results section, PARAFAC on the Del.icio.us dataset led to a category system formed of: *Music, Automotive, Programming Languages, Photo, Geography, Fashion, HowTo, Innovation.* Similarly, the categories in Citeulike were also high-quality: *Complex Adaptive Systems, Computer Security, Educa-*

tion/Pedagogy , Chemistry, Primates, Alternative Medicine, Climate Change, Mice experiments, Neuropsychology, Orthopedics. Each of the categories contains the relevant vocabularies with the added benefits of being both actual and democratic. In a sense, the vocabularies represent the mind-share of advanced Web users. The ontology can be extended simply by increasing the number of clusters.

5.5.3 Relevance of the results for the task of community identification

We believe that the task of community identification from folksonomies benefits the most from applying soft clustering methods. As previously observed, the visual representations of the three-dimensional clusters show that the same users group themselves in different clusters by their vocabularies and their interests. As an example, the user */radudragusin/* is the first in the cluster *Programming Languages* and the third in the cluster *Geo/Travel.* From the point of view of the purity measure of a community, most of the clusters obtained in the best performing 10-clusters experiments were above a purity of 0.5.

CHAPTER 6. SUMMARY AND DISCUSSION.

6.1 Related work

6.1.1 Related work on clustering and ranking in Folksonomies

Given the inherent unstructured nature of Folksonomies, clustering is one of the first Machine Learning applications that was proposed.

In Brooks and Montanez. (2006), the authors grouped articles tagged using Technocrati (http://technorati.com) by representing them in a ”bags of words” representation using their tags, weighted by TFIDF. Their findings were that the clusters were broad categories and were necessarily effective in describing the articles content. By comparison, the clusters obtained using multi-dimensional clustering are exact enough to be used for general recommendation. The authors also used agglomerative clustering to construct a hierarchy of tags.

Another clustering approach belongs to Satoshi Niwa and Honiden (2007) and it is again aimed at extracting clusters of tags. The clustering method uses in an innovative way both the object-based tag co-occurrence rate and the user-based tag co-occurrence rate in order to discover synonyms. The intuition behind the method is that the difference in object and user-based co-occurrence rate is evidence for synonymy since different users can use different tags with the same meaning , but the same user will use the same tag across similar objects.

As stated before, spectral methods on a two-dimensional representation of folksonomies have been already pursued: Begelman et al. (2006) uses the spectral bisection as in Pothen et al. (1990) (a variant of spectral clustering) algorithm to split the graph into two clusters. It recursively compares the value of the modularity function Q_0 of the original unpartitioned graph to the value of the modularity function Q_1 of the partitioned graph. If $Q_1 > Q_0$ the algorithm partitions the data and keeps recursing, otherwise it stops and returns the current partitioning.

Aside from clustering, several papers explored the idea of ranking in Folksonomies by extending classical work such as HITS (Kleinberg (1999)) and PageRank (Page et al. (1998)) to folksonomies:

Wu et al. (2006) presents a modified HITS algorithm on the graph in order to obtain experts (hubs) and high-quality documents (authorities) for a given keyword, based on the tag information.

Andreas Hotho (2006) represent folksonomies as a tripartite graph and use a modified PageRank algorithm to obtain rankings for users, tags and webpages.

6.1.2 Related work on using high-dimensional methods for clustering.

Related work using PARAFAC: TOPHITS. One of the closest approaches that uses PARAFAC is TOPHITS (Kolda et al. (2005); Kolda and Bader (2006)): In this paper the authors extend the classical HITS algorithm (Kleinberg (1999)) to include the link captions. The three dimensions are respectively source page, destination page and link caption. Each of the components in the PARAFAC decomposition has similarly to the bi-dimensional counterpart the meaning of respectively, hub, authority, and term scores for the dominant topic (or grouping) in the web page collection. The reported results are only qualitative but still compelling.

Related work using TUCKER3: CubeSVD. In Sun et al. (2005) decomposes three-dimensional click-through data in the form of ¡user,query,page¿ generated by users when searching and then navigating through the results using the MSN search engine. The results in personalizing the ranking of the search results given a user and a query were encouraging.

Related work using the Hypergraph Laplacian: Spectral clustering is a performant method, so the extension of spectral clustering to hypergraphs attracted a lot of interesting work: (Zhou et al. (2005)),(Li and Solé (1996)), (Rödl and Winkler (1989)). However, in their paper, Agarwal et al. (2006) the authors proved that the various proposals for the Hypergraph Laplacian are equivalent with two hypergraph expansions which under non-restrictive conditions are also equivalent two each other. The hypergraph expansions are the *Graph clique* expansion and the *Graph star* expansion (Zien et al. (1996)). The *Graph clique* expansion associates each hyperdge in the original graph with a clique in the extended graph and the *Graph star* expansion introduces a new vertex for each hyperedge in the original graph and connects all the vertices in the original hypergraph to their corresponding hyperedge vertices resulting in a bipartite graph. The original multi-dimensional objective is optimized by standard spectral clustering applied on the resulting graphs. We performed preliminary experiments using the method, but the lack of relevant results made us not to pursue it any further.

6.2 Summary

Folksonomies are an important source of metadata, with attractive characteristics such as democracy and actuality. It is therefore important to develop performant methods to explore and exploit them. Due to their natural three-dimensional representation they are ideal candidates for multi-dimensional decomposition methods such as PARAFAC and Tucker3. In our current work we explored their performance as clustering methods using various interpretations of their results which were inspired by their two-dimensional counterparts. The experimental results were encouraging and could be used for both Ontology Extraction and Community Identification. During the experiments we also discovered that the dimensions are not equally informative and that the clusters are naturally overlapping. In conclusion, as contributions, we introduced models and methods that allow for direct clusterization in the natural representation of the folksonomies, we compared various methods of deriving clustering indicators from the decompositions methods, we developed a quantitative method that can compare two and multi-dimensional clustering methods, we introduced a new pre-processing method and finally we showed improvement over the classical techniques.

6.3 Further work

The use of 3-dimensional decompositions is not limited to clustering and can be applied to the other reviewed tasks such as high-dimensional information retrieval and collaborative filtering of folksonomies. Also, additional dimensions can be added: a bag of words dimension representing a website, a color histogram representing an image, and so on, depending on the nature of the particular folksonomy.

APPENDIX.1 Del.icio.us qualitative results: Parafac on the Laplacian Tensor

Cluster nr.1: ***Music***

- Tags: device, portable, zune, player, microsoft, windows, mp3, audio...
- Users: /toni23polster/, /bobmah/, /pricecs/, lanzbulldog, /clickykbd/...
- Articles: zune-media-device.com/2007/02/22/ open-question-is the-zune-wall-charger-120v-or-240v does-it-matter-if-im-in-australia/, zune-media-device.com/2007/02/23/ open-question-what to-do-if-your-zune-is-stolen/, zune-media-device.com/2007/02/24/ open-question does-anyone-have-zunereactor-2/, zune-media-device.com/2007/02/24/ open-questionwouldnt-it-be-nice/, innovationzen.com/, open.bbc.co.uk/labs/, web.mit.edu/invent/h-main.html, reality.media.mit.edu/...

Cluster nr.2: ***Auto***

- Tags: auto, automobile, car, motorcycle, of, garmin, interest, traffic...
- Users: /donnachaidh/, /negativsteve/, /eddyc123/, /linlindsay/, /allume/, /kassiopea/, /joshgesler/, /kiretsu/...
- Articles: www.gps-poi-us.com/, www.radarfalle.de/ software/garmin.php, www.poihandler.com/, www.poiedit.com/, www.gpspassion.com/fr/ downloads.asp, www.scdb.info/en/, www.fartsboks.com/, www.geotourguide.com/...

Cluster nr.3: *Programming Languages*

- Tags: delphi, perl, python, basic, lisp, java, sql, c#...
- Users: /radudragusin/, /adamant1988/, /aze/, /hhoomm/, /drsalonen/, /mrm/, /seamot/, /donnachaidh/...
- Articles: www.engin.umd.umich.edu/cis/course.des/cis400/, www.freeprogrammingresources.com/, www.computer-books.us/, www.techtoolblog.com/archives/ 195-free-online programming-books, www.techbooksforfree.com/, www.webdesign.org/, www.tiobe.com/tpci.htm, www.freetechbooks.com/...

Cluster nr.4: *PhotoBlogging*

- Tags: dansk, blogging, danmark, foto, wordpress, photoblog, personal, blogger...
- Users: /nyholm/, /depmedia/, /dave9191/, /elev3n/, /slagerst/, /retrocoli/, /hypeway/, /beetle0042000/...
- Articles: photokult.net/, wiphey.com/, imagescph.dk/, johannes.jarolim.com/blog/wordpress/ yet-another-photoblog/ download-installation/ sbpages, www.fugleognatur.dk/, orderedlist.com/, www.cameraontheroad.com/, oschlag.dk/weblog/...

Cluster nr.5: *Geo/Travel*

- Tags: mapping, geography, gps, maps, map, garmin, auto, automobile...
- Users: : /donnachaidh/, /negativsteve/, /radudragusin/, /eddyc123/, /jonhillier/, /robbytherobot/, /poeticwax/, /wmplreference/...
- Articles: mywonderfulworld.org/games.html, earthobservatory.nasa.gov/newsroom/bluemarble/, www.travelbygps.com/, www.geotourguide.com/, www.radarfalle.de/software/garmin.php, www.gpspassion.com/ fr/downloads.asp, www.poihandler.com/, www.gps-poi-us.com/...

Cluster nr.6: *Fashionable*

- Tags: ummm, women, comics, cool, film, photography, clothes, art...

- Users: /robbytherobot/, /esamon/, /sunkentheory/, /allume/, /negativsteve/, /skullforce/, /clickykbd/, /anodyne99/...

- Articles: fogonazos.blogspot.com/2006/12/ la-ertica-del-robot26.html, www.ctv.es/ users/fjvidal/ archives.htm, www.tshirthell.com/ hell.shtml, www.neatorama.com/ 2007/01/02/ 13-photographs-that-changed-the-world/, www.globalorgasm.org/, files.kavefish.com/ pictures/collections/ pictures-from-the-sky/ index-list.html, gigapedia.org/, www.oculture.com/weblog/ 2006/10/audio-book-podc.html...

Cluster nr.7: *Directory*

- Tags: iyp, pages, yellow, yellowpages, directory, search, maps, phone...

- Users: /lanzbulldog/, /crfarnum/, /donturn/, /klabol/, /rickbissonnette/, /matsch-o0/, /lazlo101/, /bergerx/...

- Articles: switchboard.com/, www.truelocal.com/, www.superpages.com/, www.sensis.com.au/, www.yellowpages.com/, www.citysearch.com/, www2.metrobot.com/ def.cfm?j=1, www.judysbook.com/...

Cluster nr.8: *HowTo/Community*

- Tags: guide, lists, sharing, songs, favorites, guides, knowledge, reviews...

- Users: /linlindsay/, /allume/, /joshgesler/, /lisku/, /kiretsu/, /blue-ocean/, /ocoiso/, /pauloe/...

- Articles: www.listal.com/, www.gnoosic.com/, www.riffs.com/, www.fiql.com/, www.etsy.com/, www.econsultant.com/, answers.yahoo.com/, www.listology.com/index.cfm...

Cluster nr.9: *Innovation*

- Tags: iswappy, innovation, research, ideas, networks, business, information, mit...

- Users:/bobmah/, /lr1/, /oniwe/, /designlab.no/, /pricecs/, /slagerst/, /daniele85/, /businessorati/...

- Articles: web.mit.edu/invent/h-main.html, open.bbc.co.uk/labs/, innovationzen.com/, reality.media.mit.edu/, www.networkworld.com/ community/?q=node/10959, slashdot.org/ article.pl?sid=07/01/02/237223, www.seomoz.org/blog/ 21-tactics-to-increase-blog-traffic, www.briansolis.com/...

Cluster nr.10: *Web2.0*

- Tags: webapps, services, webservices, mashups, backup, archive, utilities, api...

- Users: /z303/, /cbgreenwood/, /klabol/, /mzn/, /mountchuck/, /leebax/, /malheiro/, /joshgesler/...

- Articles: www.quickonlinetips.com/ archives/2005/03/ great-flickr-tools-collection/, jeremy.zawodny.com/ blog/archives/007641.html, pipes.yahoo.com/, radar.oreilly.com/ archives/2007/02/pipes-and-filte.html, www.imified.com/, www.fuckedgoogle.com/, plagger.org/, soundmoneytips.com/article/25854...

APPENDIX.2 Del.icio.us qualitative results: SVD on the UT Laplacian matrix

Cluster nr.1: ***?***

- Tags: symbian, ummm, to, nickles, install, newspaper, virus, location...
- Users: /3un/, /adamant1988/, /adultdatingdr1/, /albeza/, /alegraphics00/, /allume/, /ambersatt/, /anodyne99/...
- Articles: www.blomsterkbh.dk/, www.befreite-dokumente.de/, www.heise.de/ security/dienste/ antivirus/ massnahmen.shtml, www.heise.de/ newsticker/meldung/85742/from/atom10, www.nickles.de/ static-cache/538158011.html, topsexmovies.blogspot.com/? tag=sexhardcore, www.todolomovil.com/, tinnus.gp32z.com/ljp/...

Cluster nr.2: ***WebDesign***

- Tags: design, css, webdesign, web, blog, reference, tools, web2.0...
- Users: /3un/, /adamant1988/, /alegraphics00/, /allume/, /ambersatt/, /anodyne99/, /antichris/, /archimac/...
- Articles: www.lifehacker.com/, www.huddletogether.com/ projects/lightbox2/, typetester.maratz.com/, www.smashingmagazine.com/ 2007/01/19/ 53-css-techniques you-couldnt-live-without/, www.smashingmagazine.com/ 2007/02/09/83-beautiful-wordpress-themes-you-probably-havent-seen/, www.webdesignfromscratch.com/ web-2.0-design-style-guide.cfm, wordpress.org/...

Cluster nr.3: *WebDesign*

- Tags: design, css, webdesign, web, inspiration, programming, blog, reference...

- Users: /3un/, /adamant1988/, /adultdatingdr1/, /albeza/, /alegraphics00/, /allume/, /ambersatt/, /anodyne99/...

- Articles:www.smashingmagazine.com/ 2007/02/09/ 83-beautiful-wordpress themes-you-probably-havent-seen/, www.smashingmagazine.com/ 2007/01/19 /53-css-techniques you-couldnt-live-without/, www.dynamicdrive.com/, www.freeprogrammingresources.com/, www.webdesignfromscratch.com/ web-2.0-design style-guide.cfm, typetester.maratz.com/...

Cluster nr.4: *Media*

- Tags: music, media, audio, video, mp3, search, social, tv...

- Users: /3un/, /adamant1988/, /adultdatingdr1/, /allume/, /ambersatt/, /anodyne99/, /antichris/, /archimac/...

- Articles: btjunkie.org/, zune-media-device.com/ 2007/02/23/ open-question what-to-do-if-your-zune-is-stolen/, zune-media-device.com/ 2007/02/24/ open-question wouldnt-it-be-nice/, zune-media-device.com/ 2007/02/24/ open-question does-anyone-have-zunereactor-2/, zune-media-device.com/ 2007/02/22/ open-question is-the-zune-wall-charger-120v-or-240v does-it-matter-if-im-in-australia/, www.listal.com/...

Cluster nr.5: *OpenSource*

- Tags: linux, programming, opensource, software, php, python, perl, mobile...

- Users: /3un/, /adamant1988/, /allume/, /ambersatt/, /anodyne99/, /antichris/, /archimac/, /ashworth102680/...

- Articles: www.freeprogrammingresources.com/, www.techtoolblog.com/archives/ 195-free-online programming-books, www.computer-books.us/, www.fring.com/, www.engin.umd.umich.edu/cis/ course.des/cis400/, www.openmoko.com/...

Cluster nr.6: ***Shopping***

- Tags: shopping, recipes, deals, bargains, food, cooking, shop, sell...

- Users: /3un/, /adamant1988/, /adultdatingdr1/, /allume/, /ambersatt/, /anodyne99/, /antichris/, /archimac/...

- Articles: www.rvgnet.net/used-games/, www.gamepawn.com/index.php?session=, www.fatwallet.com/, allrecipes.com/, www.hot-deals.org/, www.galison.com/index.aspx, www.foodnetwork.com/, frugalcuisine.blogspot.com/...

Cluster nr.7: ***Music***

- Tags: music, audio, programming, mp3, drumnbass, dnb, device, portable...

- Users: /3un/, /adamant1988/, /adultdatingdr1/, /alegraphics00/, /allume/, /ambersatt/, /anodyne99/, /antichris/...

- Articles: www.dogsonacid.com/forumdisplay.php?forumid=4, www.freeprogrammingresources.com/, www.dogsonacid.com/, zune-media-device.com/ 2007/02/23/ open-question what-to-do-if-your-zune-is-stolen/, zune-media-device.com/ 2007/02/24/ open-question wouldnt-it-be-nice/, zune-media-device.com/ 2007/02/24/ open-question does-anyone-have-zunereactor-2/, zune-media-device.com/ 2007/02/22/ open-question is-the-zune-wall-charger 120v-or-240v-does-it-matter if-im-in-australia/, www.breakbeat.co.uk/...

Cluster nr.8: ***Phone***

- Tags: voip, mobile, phone, skype, symbian, site, singles, pages...

- Users: /3un/, /adamant1988/, /adultdatingdr1/, /alegraphics00/, /allume/, /ambersatt/, /anodyne99/, /antichris/...

- Articles: www.fring.com/, www.nimbuzz.com/en/, www.peerme.com/en/index.php, switchboard.com/, www.openmoko.com/, www.yellowpages.com/, btjunkie.org/, www.superpages.com/...

Cluster nr.9: *Modern Living*

- Tags: recipes, cooking, food, career, finance, job, health, jobs...
- Users: /adamant1988/, /adultdatingdr1/, /allume/, /ambersatt/, /anodyne99/, /archimac/, /ashworth102680/, /aze/...
- Articles: allrecipes.com/, www.foodnetwork.com/, frugalcuisine.blogspot.com/, www.carbohydrate-guide.com/archives/2005/09/16/thermogenic-foods.html, www.rockportinstitute.com/resumes.html, soundmoneytips.com/article/25854, creditpro.wordpress.com/ 2007/02/11/ 8-things-you-must-do if-your-identity-is-stolen/, extratasty.com/...

Cluster nr.10: *WebDesign*

- Tags: css, inspiration, design, webdesign, flash, typography, color, usability...
- Users: /3un/, /adamant1988/, /adultdatingdr1/, /alegraphics00/, /allume/, /ambersatt/, /anodyne99/, /antichris/...
- Articles: typetester.maratz.com/, kuler.adobe.com/, www.thefwa.com/, roxik.com/pictaps/, www.smashingmagazine.com/2006/11/11/ css-based-forms-modern-solutions/, www.truelocal.com/, www.dontclick.it/, btjunkie.org/...

APPENDIX.3 Citeulike qualitative results: Parafac on the Laplacian Tensor

Cluster nr.1: ***Complex Adaptive Systems***

- Tags: ant-colony-systems, collective-systems, adaptive-computation, distributed-computing, collective-computing, artificial-life, stigmergy, complex-systems...

- Articles: ”Artificial Ant Colonies in Digital Image Habitats - A Mass Behaviour Effect Study on Pattern Recognition”, ”Varying the Population Size of Artificial Foraging Swarms on Time Varying Landscapes”, ”On Ants, Bacteria and Dynamic Environments”, ”Societal Implicit Memory and his Speed on Tracking Extrema over Dynamic Environments using Self-Regulatory Swarms”, ”Evolving A Stigmergic Self Organized Data Mining”...

Cluster nr.2: ***Computer Security***

- Tags: spyware, malware, rid, manually, uninstall, remove, delete, how,...

- Articles: ”411 Spyware Report : Ultimate Defender Removal Instructions”, ”411 Spyware Report : Pest Trap Removal Instructions”, ”411 Spyware Report : SafeAndClean Removal Instructions”, ”411 Spyware Report : VirusBurster Removal Instructions”, ”411 Spyware Report : Adware Punisher Removal Instructions”...

Cluster nr.3: ***Education/Pedagogy***

- Tags: committee, council, board, advisory, gs, employers, advisors, post-secondary...

- Articles: ”Project EMD-MLR: Educational Material Development and Research in Machine Learning for Undergraduate Students”, ”How institutions respond to Training Packages”, ”Curriculum restructure to answer critical needs in packaging for energy efficiency/renewable energy systems, wireless, and mixed-signalsystems areas”,”A contemporary review of tourism degrees in the united kingdom”...

Cluster nr.4: *Chemistry*

- Tags: period-osu, he, waals, der, van, co, ar, potential...
- Articles: ”Structure and energetics of van der Waals complexes of carbon monoxide with rare gases. He-CO and Ar-CO”, ”Intermolecular forces via hybrid Hartree-Fock plus damped dispersion (HFD) energy calculations. Systems with small nonsphericity: Ar-H2, Ne-H2, and He-H2”, ”A new He-CO interaction energy surface with vibrational coordinate dependence. I. Ab initio potential and infrared spectrum”, ”Ab initio potential energy surface and dynamics of He-CO”, ”Anisotropic intermolecular forces. I. Rare gas-hydrogen chloride systems”...

Cluster nr.5: *Primates*

- Tags: great-apes, gorillas, chimpanzees, lope, gabon, monkeys, diet, frugivory...
- Articles: ”Composition of the Diet of Chimpanzees and Comparisons with That of Sympatric Lowland Gorillas in the Lope Reserve, Gabon”, ”Ecology and social organisation of African rainforest primates: relevance for understanding the transmission of retroviruses”, ”The primate community of the Lop Reserve, Gabon: Diets, responses to fruit scarcity, and effects on biomass”, ”Seed dispersal by a diurnal primate community in the Dja Reserve, Cameroon”,”Seasonal variation in the feeding ecology of the grey-cheeked mangabey (Lophocebus albigena) in Cameroon”...

Cluster nr.6: *Alt. medicine*

- Tags: herbal, liquid, pressure, drugschinese, chromatographyhigh, chemistry, quality, phytopharmaceutical...

- Articles: ”Studies on quality control standard of zhishidaozhi tabloid pills”, ”Studies on quality control of Xueyakang capsule”, ”Determination of 2,3,5,4’-tetrahydroxystilbene-2-O-beta-D-glucoside in yangan oral liquid by HPLC”, ”Studies on quality control standard of zhishidaozhi tabloid pills”,”Study on quality control of effective fraction in qixue bingzhi decoction”...

Cluster nr.7: ***Climate change***

- Tags: fif, climate-change, forest, climate, forests, co2, soil, vegetation...

- Articles: ”The effects of climate change on decomposition processes in grassland and coniferous forests”, ”The CO2 dependence of photosynthesis, plant growth responses to elevated CO2 and their interaction with soil nutrient status, II. Temperate and boreal forest productivity and the combined effects of increasing CO2 and increased nitrogen deposition at a global scale”, ”The use of iron and other trace-element fertilizers in mitigating global warming”, ”Responses in NPP and carbon stores of the northern biomes to a CO2 induced climatic-change, as evaluated by the Frankfurt biosphere model (FBM)”...

Cluster nr.8: ***Mice experiments***

- Tags: endnote, lupus, erythematosus, inbred, mice, mrl, lpr, govt...

- Articles: ”Effects of FTY720 in MRL-lpr/lpr mice: therapeutic potential in systemic lupus erythematosu”, ”B cell anergy and systemic lupus erythematosus”, ”Increased plasma cell frequency and accumulation of abnormal syndecan-1plus T-cells in Igmu-deficient/lpr mice”, ”Toll-like receptors and activation of autoreactive B cells”, ”Monocytosis and accelerated activation of lymphocytes in C1q-deficient autoimmune-prone mice”...

Cluster nr.9: *Neuropsychology*

- Tags: ma-biblio-mehdi-juin07c, animals, neuronsphysiology, rats, prefrontal, ma-biblio-mehdi-juin07a, conditioning, learningphysiology...

- Articles: ”Involvement of basal ganglia and orbitofrontal cortex in goal-directed behavior”, ”Limbic cortical-ventral striatal systems underlying appetitive conditioning”, ”Neuronal responses in the ventral striatum of the behaving macaque”, ”Medial prefrontal cortex cells show dynamic modulation with the hippocampal theta rhythm dependent on behavior”, ”Modeling functions of striatal dopamine modulation in learning and planning”...

Cluster nr.10: *Orthopedic*

- Tags: orth-main, bone, cartilage, tissue, mechanical, collagen, human, joint...

- Articles: ”Biomechanical characterization and in vitro mechanical injury of elderly human femoral head cartilage: comparison to adult bovine humeral head cartilage”, ”Osteoarthritis and osteoporosis: clinical and research evidence of inverse relationship”, ”Compressive properties of mouse articular cartilage determined in a novel micro-indentation test method and biphasic finite element model”, ”Age variations in the properties of human tibial trabecular bone and cartilage”, ”Fibroblast growth factor (FGF) 18 signals through FGF receptor 3 to promote chondrogenesis”,”Blocking the effects of IL-1 in rheumatoid arthritis protects bone and cartilage”...

APPENDIX.4 Citeulike qualitative results: SVD on the UT Laplacian matrix

Cluster nr.1: *Genetics?*

- Tags: key-gene-expression, gaze, emodin, epimedium, link-mining, pueraria, quinolizines...
- Articles: ”Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments”, ”A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression”, ”Evolutionary changes in cis and trans gene regulation”, ”Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae”...

Cluster nr.2: *Genetics?*

- Tags: key-gene-expression, bibtex-import, support, govt, animals, humans, non-us, research...
- Articles: ”The metabolic world of Escherichia coli is not small.”, ”Reverse engineering of biological complexity”, ”Comparative assessment of large-scale data sets of protein-protein interactions”,”Random graphs with arbitrary degree distributions and their applications”, ”The evolutionary origin of complex features”, ”Organization, development and function of complex brain networks”

Cluster nr.3: *Genetics?*

- Tags: bibtex-import, support, govt, animals, non-us, humans, male, us...

- Articles: The metabolic world of Escherichia coli is not small.”, ”Reverse engineering of biological complexity”, ”Comparative assessment of large-scale data sets of protein-protein interactions”,”Motifs in Brain Networks”, ”The evolutionary origin of complex features”, ”Organization, development and function of complex brain networks”...

Cluster nr.4: ***Computer/Network security?***

- Tags: networktheory, socialnetworks, spyware, uninstall, remove, malware, manually, rid...

- Articles: The metabolic world of Escherichia coli is not small.”, ”Reverse engineering of biological complexity”, ”Comparative assessment of large-scale data sets of protein-protein interactions”,”Motifs in Brain Networks”, ”The evolutionary origin of complex features”, ”Exploring complex networks”...

Cluster nr.5: ***?***

- Tags: networktheory, socialnetworks, fif, animals, climate-change, govt, support, climate...

- Articles: ”Reverse engineering of biological complexity”, ”Exploring complex networks”, ”Comparative assessment of large-scale data sets of protein-protein interactions”, ”Collective dynamics of ’small-world’ networks”, ”Network motifs: simple building blocks of complex networks.”

Cluster nr.6: ***Electronics***

- Tags: macroelectronics, flexible-electronics, thin-metal-film, polymer-substrate, platform, mechanics, localization, adhesion...

- Articles: ”Reverse engineering of biological complexity”, ”Exploring complex networks”, ”Comparative assessment of large-scale data sets of protein-protein interactions”, ”The evolutionary origin of complex features”, ”Organization, development and function of complex brain networks”

Cluster nr.7: *Administration*

- Tags: council, committee, gs, board, advisory, employers, 1996, advisors...
- Articles: ”Reverse engineering of biological complexity”,”Exploring complex networks”, ”Comparative assessment of large-scale data sets of protein-protein interactions”, ”Random graphs with arbitrary degree distributions and their applications”...

Cluster nr.8: *Folksonomies*

- Tags: tagging, folksonomy, no-tag, collaboration, web, networks, network, social...
- Articles: ”Reverse engineering of biological complexity”,”Exploring complex networks”, ”Comparative assessment of large-scale data sets of protein-protein interactions”, ”Random graphs with arbitrary degree distributions and their applications”...

Cluster nr.9: *Folksonomies*

- Tags: tagging, folksonomy, no-tag, collaboration, networks, support, govt, network...
- Articles: ”Reverse engineering of biological complexity”,”Exploring complex networks”, ”Comparative assessment of large-scale data sets of protein-protein interactions”, ”Random graphs with arbitrary degree distributions and their applications”, ”Organization, development and function of complex brain networks”...

Cluster nr.10: *Complex Adaptive Systems*

- Tags: collective-systems, adaptive-computation, ant-colony-systems, collective-computing, distributed-computing, artificial-life, swarm-intelligence, artificial-intelligence...
- Articles: ”Reverse engineering of biological complexity”,”Exploring complex networks”,”The anatomy of a large-scale hypertextual Web search engine”, ”Classes of small-world networks”, ”Finding and evaluating community structure in networks”

BIBLIOGRAPHY

Agarwal, S., Branson, K., and Belongie, S. (2006). Higher order learning with graphs. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 17–24, New York, NY, USA. ACM Press.

Andreas Hotho, R. J. (2006). Folkrank: A ranking algorithm for folksonomies.

Ausubel, D., Novak, J., and Anesian, H. (1978). *Educational Psychology: A cognitive view*. Holt, Rinehart, and Winston, 2 edition.

Baader, F. (2003). *The Description Logic Handbook : Theory, Implementation and Applications*. Cambridge University Press.

Bader, B. W. and Kolda, T. G. (2007). Efficient MATLAB computations with sparse and factored tensors. *SIAM Journal on Scientific Computing*, 30(1):205–231.

Bao, J., Caragea, D., and Honavar, V. (2006). Towards collaborative environments for ontology construction and sharing. In Mcquay, W. K. and Smari, W. W., editors, *International Symposium on Collaborative Technologies and Systems (CTS'06)*, pages 99–108.

Begelman, G., Keller, P., and Smadja, F. (2006). Automated tag clustering: Improving search and exploration in the tag space.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web: Scientific american. *Scientific American*.

Brooks, C. H. and Montanez., N. (2006). Improved annotation of the blogopshere via autotagging and hierarchical clustering. In *WWW06: Proceedings of the 15th World Wide Web Conference*, Edinburgh, Scotland. ACM Press.

Carroll, J. and Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition. *Psychometrika*, 35(3):283–319.

Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *Knowledge Discovery and Data Mining*, pages 269–274.

Dhillon, I. S., Guan, Y., and Kulis, B. (2004). Kernel k-means: spectral clustering and normalized cuts. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–556, New York, NY, USA. ACM Press.

Ding, C. and He, X. (2004). ¡i¿k¡/i¿-means clustering via principal component analysis. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, New York, NY, USA. ACM.

Editorial (2005). Wiki's wild world.

Ganter, B. and Wille, R. (1998). *Formal Concept Analysis: Mathematical Foundations*. Springer.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*. The Johns Hopkins University Press, third edition edition.

Harshman, R. A. and Lundy, M. E. (1994). Parafac: Parallel factor analysis. *Computational Statistics & Data Analysis*, 18(1):39–72.

Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. *Applied Statistics*, 28:100–108.

Heymann, P. and Garcia-Molina, H. (2006). Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Stanford University.

Kang, D.-K., Silvescu, A., Zhang, J., and Honavar, V. (2004). Generation of attribute value taxonomies from data for data-driven construction of accurate and compact classifiers. In *ICDM '04: Proceedings of the Fourth IEEE International Conference on Data Mining*, pages 130–137, Washington, DC, USA. IEEE Computer Society.

Kiers, H. A. L. and Van Mechelen, I. (2001). Three-way component analysis: Principles and illustrative application. *Psychological Methods*, 6:84–110.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

Kolda, T. and Bader, B. (2006). The TOPHITS model for higher-order web link analysis. In *Workshop on Link Analysis, Counterterrorism and Security*.

Kolda, T. G., Bader, B. W., and Kenny, J. P. (2005). Higher-order web link analysis using multilinear algebra. In *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 242–249, Washington, DC, USA. IEEE Computer Society.

Leeuw, J., Young, F., and Takane, Y. (1976). Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 41(4):471–503.

Li, W.-C. W. and Solé, P. (1996). Spectra of regular graphs and hypergraphs and orthogonal polynomials. *Eur. J. Comb.*, 17(5):461–477.

Long, B., Wu, X., Zhang, Z. ., and Yu, P. S. (2006). Unsupervised learning on k-partite graphs. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 317–326, New York, NY, USA. ACM Press.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project.

Paolillo, J. C. and Penumarthy, S. (2007). The social structure of tagging internet video on del.icio.us. pages 85–85.

Pearson, K. (1901). On lines and planes of closest fit to points in space. *Philosophical Magazine*, 2:559–572.

Pothen, A., Simon, H., and Liou, K.-P. (1990). Partitioning sparse matrices with eigenvectors of graphs. *SIAM Journal of Matrix Analysis and Applications*, 11:430–452.

Prieur, C., Cardon, D., Beuscart, J.-S., Pissard, N., and Pons, P. (2008). The stength of weak cooperation: A case study on flickr.

Rödl, V. and Winkler, P. (1989). A Ramsey-type theorem for orderings of a graph. *j-SIAM-J-DISCR-MATH*, 2(3):402–406.

Satoshi Niwa, T. D. and Honiden, S. (2007). Folksonomy tag organization method based on the tripartite graph analysis. In *SWeCKa 2007: IJCAI Workshop on Semantic Web for Collaborative Knowledge Acquisition*, Hyderabad, India.

Schmitz, C. (2006). Small world folksonomies: Clustering in tri-partite hypergraphs.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Smilde, A. K. (1992). Three-way analyses: Problems and prospects. *Chemometrics and Intelligent Laboratory Systems*, 15:143–158.

Specia, L. and Motta, E. (2007). Integrating folksonomies with the semantic web. pages 624–639.

Sun, J.-T., Zeng, H.-J., Liu, H., Lu, Y., and Chen, Z. (2005). Cubesvd: a novel approach to personalized web search. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 382–390, New York, NY, USA. ACM Press.

Trefethen, L. N. and Bau, III, D., editors (1997). *Numerical Linear Algebra*.

Tucker, L. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311.

Vasilescu, M. A. O. and Terzopoulos, D. (2002). Multilinear analysis of image ensembles: Tensorfaces. In *ICPR(2)*, pages 511–514.

Wu, H., Zubair, M., and Maly, K. (2006). Harvesting social knowledge from folksonomies. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 111–114, New York, NY, USA. ACM Press.

Yeung, C. M. A., Gibbins, N., and Shadbolt, N. (2007). Understanding the semantics of ambiguous tags in folksonomies. In *The International Workshop on Emergent Semantics and Ontology Evolution (ESOE2007) at ISWC/ASWC 2007*.

Zhou, D., Huang, J., and Schlkopf, B. (2005). Beyond pairwise classification and clustering using hypergraphs. Technical Report 143, Tbingen, Germany.

Zien, J. Y., Schlag, M. D. F., and Chan, P. K. (1996). Multi-level spectral hypergraph partitioning with arbitrary vertex sizes. In *ICCAD '96: Proceedings of the 1996 IEEE/ACM international conference on Computer-aided design*, pages 201–204, Washington, DC, USA. IEEE Computer Society.