Hypergraph Modelling and Graph Clustering Process Applied to Co-word Analysis

Xavier Polanco, Eric Sanjuan

▶ **To cite this version:**

Xavier Polanco, Eric Sanjuan. Hypergraph Modelling and Graph Clustering Process Applied to Co-word Analysis. ISSI 2007 - 11th International Conference of the International Society for Scientometrics and Informetrics, Jun 2007, Madrid, Spain. pp.613-618. hal-00165984

HAL Id: hal-00165984
https://hal.science/hal-00165984

Submitted on 30 Jul 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

11th International Conference of the International Society for Scientometrics and Informetrics, Madrid, Spain, June 25-27, 2007. Proceedings, vol. 2, p. 613-618

Hypergraph Modelling and Graph Clustering Process Applied to Co-word Analysis

Xavier Polanco* and Eric Sanjuan **

*xavier.polanco@lip6.fr
LIP6, Université Pierre et Marie Curie, CNRS UMR 7606,
104 avenue du Pte. Kennedy, 75016 Paris, France
**eric.sanjuan@iniv-avignon.fr
LIA, Université d'Avignon, 339 chemins des Meinajaries, 84911Avigon, France

Abstract
We argue that any document set can be modelled as a hypergraph, and we apply a graph clustering process as a way of analysis. A variant of the single link clustering is presented, and we assert that it is better suited to extract interesting clusters formed along easily interpretable paths of associated items than algorithms based on detecting high density regions. We propose a methodology that involves the extraction of similarity graphs from the indexed-dataset represented as a hypergraph. The mining of informative short paths or geodesics in the graphs follows a graph reduction process. An application for testing this methodology is briefly exposed. We close this paper indicating the future work.

Keywords
Co-word analysis, graph clustering, hypergraph modelling.

1. Introduction

The main objective of this paper is to propose a graph clustering methodology including hypergraph modelling of raw data. The position that we support here can be summarized in these three proposals: [1] the hypergraph modelling constitutes a uniform data model that allows us to generalize the method of co-word analysis as we try in this paper, and it can be compared to Galois lattice modelling use by formal concept analysis (FCA) (Wille, 1982). This last point we will not treat it here. [2] On the basis of what has been observed in co-word analysis (Courtial, 1990), short paths of strong associations can reveal potential new connections between separated sectors of the co-word network. Thus, the graph clustering method that we propose does not focus on homogeneous clusters, but highlight some heterogeneous clusters formed along short path of strong associations. [3] Concerning the graph clustering, we support that a variant of the single link clustering (SLC) is better suited to build interesting clusters, formed along easily interpretable paths of associated items, than algorithms based on detecting high density regions.

Section 2 introduces the hypergraph modelling, sections 3 and 4 deal with the graph reduction and clustering process. A real application on SCI dataset try, in section 5, to provide the empirical demonstration of the approach exposed in the former sections.

2. Hypergraph Modelling

Hypergraph definition (Berge, 1987): A hypergraph is a pair (X, E) where X is a set of elements, called hypernodes or hypervertices, and E is a set of hyperedges; a hyperedge e is a non empty subset of X, $S(X)\backslash\emptyset$. While graph edges are pairs of nodes, hyperedges are arbitrary set of nodes or vertices, because in a hypergraph the edges can connect any number of vertices. A hypergraph is a family of sets drawn from the set X.

Any set of documents D can be modelled as a hypergraph H. Each of the documents can be represented by a hypervertex, and the document elements as for instance authors, keywords, and citations are represented each one by the subset of documents sharing these elements. These subsets

constitute the hyperedges of the hypergraph. In the sequel, we shall identify each keyword w with the subset of documents $\{d_1, \ldots, d_p\}$ indexed by w. According to the type of information that we want to analyse, intersection graphs that we shall call co-occurrence graphs, can be derived from different subparts of H. The vertices V are the selected hyperedges meanwhile an edge is drawn among two vertices (v_i,v_j) whenever they have a non empty intersection. For example, if we select as subpart of H, the hypervertex representing keywords or authors or citations, the resulting intersection graph is the keyword or co-author or co-citation graph. If we select both author and keyword hyperedges, we obtain a graph of associations between keywords and authors; if we select both citation and keyword hyperedges, we obtain a graph of associations among keywords and citations.

3. Graph reduction process

The next consists of applying a graph clustering process from the co-occurrence data matrix, where the data can be keywords, authors or citations. Here, we limit to treat the co-word matrix. This is the input matrix of the graph of co-occurrences $Go(V,E)$.

Many similarities between keywords can be defined based on the cardinality of $w_i \cap w_j$ (Van Cutsem, 1994), in the present case the intersection graph $Go(V,E)$ of H defines the co-word graph having as many edges as there are non null values in the similarity matrix. The set E of edges of $Go(V,E)$ is the set of pairs of keywords $\{w_i,w_j\}$ where $w_i \cap w_j \neq \emptyset$. But the associations between keywords cannot be considered in a crisp binary way. The co-occurrence frequency alone is not enough to measure the strength of associations, because it favours high-frequency couples compared to those with low frequency. Then, it is necessary to use some normalizing coefficient. For this task, we apply the "equivalence coefficient" (as originally defined in Michelet, 1988) based on the product of conditional probabilities of appearance of a term knowing the presence of the other one. The equivalence coefficient that we note $\sigma(w_i,w_j) = |w_i \cap w_j|^2 / (|w_i| \times |w_j|)$ allows to avoid weak relations and to normalise frequency of keywords. This coefficient also has an easy interpretation in terms of probability theory, since it is the product of conditional probabilities of finding one item knowing the presence of the other. This coefficient is maximized by pairs of items that are in the same closed sets. We denote $G(A) = (V,E,\sigma)$ the weighted graph of associations. Usually, when a co-occurrence matrix is used, a threshold is set on the keyword frequency in order to obtain a less sparse matrix. On the other hand, setting the threshold on association value $\sigma(w_i,w_j)$ and not only on keyword frequency $|w_i|$ is better suited to form clusters that are closed. Consequently, every value s in $]0,1[$ induces a sub-graph $G(A>s) = (V,E_s,\sigma)$ where E_s is the set of pairs of vertices (w_i,w_j) such that $\sigma(i,j) > s$.

4. Graph clustering

We apply a variant algorithm of the single link clustering (SLC) called CPCL (Classification by Preferential Clustered Link), originally introduced in Ibekwe-SanJuan (1998), here we use its optimised version (SanJuan et al, 2005). This algorithm is applied on the graph $G(A>s)$ building clusters of keywords related by geodesic paths that are constituted of relatively high associations. Any variant of single link clustering that reduces its chain effect can produce interesting results in this context, since they naturally form clusters along short geodesics of maximal weight. The CPCL algorithm (see table 1) merges iteratively clusters of keywords related by an association strongest than any other in the external neighbourhood. In other words, CPCL works on local maximal edges instead of absolute maximal values like the standard SLC. We refer the reader to Berry et al. (2004) for a detailed description of the algorithm in the graph formalism.

Table 1 CPCL Algorithm

Program CPCL (V,E, σ)

1) Compute the set S of edges {i, j}such that σ(i,j) is greater than s(i,z), and s(j,z) for any vertex z
2) Compute the set C of connected components of the sub-graph (V,S)
3) Compute the reduced valued graph (C, E_C, σ_C), where E_C is the set of pairs of components {I,J} such that there exists {i,j} in E with i in J, j in J, and σ_C(I,J) = max{σ(i,j) : I in I, j in J).

If V <> C go to phase 1 else return (C, E_C, σ_C)

5. Experimentation

We test this graph clustering process on a corpus of 5,795 bibliographic data on "data and text mining", extracted from the SCI database, over period 2000-2006. The different components of the bibliographic data are catalogued according to 39 fields coded by a combination of two capitals as specified by SCI. These 5,795 data characterized by these 39 fields are submitted to hypergraph modelling. The hypergraph, denoted H, is stored on the form of a ternary relation i.e. records involve an identification number, a code field, and keywords. The coded fields form the vertices of the hypergraph H. From this general hypergraph with 165,850 vertices and 5,795 edges, a sub-hypergraph by selecting a subset of vertices is extracted. In this case we select the set of keywords (SCI field denoted by DE). The result is a sub-hypergraph K with 8,040 vertices and 3,171 edges, the vertices are the author keywords. The degree of hyperedges corresponds to the number of keywords indexing a reference. Its minimal value is 1, the mean 5.11 and the maximal 35. The degree of vertices is the number of references indexed by a keyword. On this corpus, its minimal value is 2, its mean 6.35 and its maximal value 1,722 for keyword "data mining". The frequency of all other keywords is lower than 200.

The hypergraph modelling allows introducing a new data analysis instrument: the minimal transversal. For any subset S of hyper-vertices, let us denote by $T(S)$ the set of hyper-edges h such that $h \cap S \neq \emptyset$, S is said to be a minimal transversal if for any s in S, $T(S - s) \neq T(S)$. Consequently, the minimal transversals correspond to closed item-sets in data mining (Zaki, 2004), and to formal concepts in Formal Concept Analysis (Ganter et al, 2005; 1999). The number of hyperedges in which S is included is the support of S. On K there are 2,526 minimal transversals with at most 5 elements and a support greater that 0.001. Table 2 gives some examples of minimal traversals with 5 elements having a largest support.

Table 2 Minimal traversals of the hypergraph K.

Support					
0.00421	bioinformatics	cancer	data mining	genomics	proteomics
0.00316	genomics	machine learning	microarray	proteomics	text mining
0.00316	data mining	dimensionality reduction	feature extraction	neural network	pattern recognition
0.00316	clustering	machine learning	microarray	proteomics	text mining
0.00316	bioinformatics	genomics	machine learning	proteomics	text mining
…	…	…	…	…	…

Computing such sets allow detecting subsets of frequently associated items. However, computing all minimal transversal is intractable since the number of such sets can be exponential on the number of items, because the numerous overlaps between minimal transversal.

The graph of weighted associations $G(A)$ constitutes the intersection graph of the hypergraph K with 8,037 vertices and 34,375 weighted edges. The $G(A)$ graph is a small world graph (SWG) since its average clustering measure is 0.47. This value is far from the expected value for a random graph having the same average degree which is the average degree over the number of edges 4.43/2,335. As in random graphs, the average path length is low 2.31. These are the two conditions usually considered to characterise SWGs (Watts, 1999). The SWGs are compact graphs with a high number of simplicial vertices that are vertices whose neighborhood forms a complete graph. Setting a low threshold on association values (0.001 here) is enough to drastically reduce the number of edges and reduces its clustering measure (0.35). However, we do not loose the SW property since 0.35 is much greater than the mean degree 3.28 over the number of edges 1,057, and the average path length is low 4.12.

The graph $G(A>s)$ is induced from $G(A)$ fixing the threshold s at 0,001, and the co-occurrence frequency at 2. The graph $G(A>0.001)$ has 447 vertices and 1,017 edges and easier to handle. $G(A>0.001)$ presents a central dense connected component of 187 vertices and 369 edges. All other components are small (less than 10% of the total number of vertices). We focus our experiment on this component to observe if it allows detecting complex associations as the one pointed out by minimal transversals. Applying the CPCL algorithm on $G(A>0.001)$ we obtain a clustered graph, G(CPCL), with 187 vertices and 738 edges, each vertex is a cluster. By definition, CPCL output tries to highlight

a disjoint family of clusters formed on geodesics of relatively high associations. We have experimentally checked that most of these clusters are minimal transversals. The biggest one has 10 keywords, and the mean size is 4.76% of these clusters are minimal transversals of the hypergraph *K*. Table 3 shows three examples of clusters of medium size which are minimal transversal.

Table 3 Clusters in *G*(CPCL) graph

Label	vertex 1	vertex 2	vertex 3	vertex 4
Ontology	thesaurus	query processing	ontology	knowledge mining
OLAP	rule based reasoning	OLAP	data interchange	XML
Intrusion detection	computer security	intrusion detection	anomaly detection	user profiling

They are labeled by the vertex having in the graph the highest betweenness valued, and the elements are enumerated following the geodesic that cross the edges with a highest value. This is the way that the clusters are extracted.

A numerical notation of the keywords (i.e. co-words) of the graph *G*(*A*>0.001), and also of the clusters of the graph *G*(CPCL), gives an idea of their positions in the structure of the graphs. The structural properties considered are centrality, density, betweenness, degree, and w_betw. Centrality (Centrl) gives the sum of association values involving the vertex, density (Dens) gives the number of edges in the neighborhood over the maximal theoretic number, betweenness (Betw) gives the number of geodesics crossing the vertex, degree gives the number of vertices in the neighborhood, and weighted betweenness (w_Betw) gives the number of geodesics that maximize the sum of association values. Tables 4 and 5 summarize the first five items ordered by decreasing betweenness (Betw). Keywords having the higher betweenness centrality point out keywords shared by the longest minimal transversals showed in table 2. This observation deserves a thorough analysis.

Table 4 Keyword vertices in *G*(*A*>0.001) graph

Vertex	Centrl	Dens	Betw	Degree	w_Betw
text mining	0.07	0.01	14,573	44	11,264
machine learning	0.08	0.03	10,243	30	12,664
bioinformatics	0.13	0.03	8,113	34	3,549
classification	0.1	0.03	7,376	23	4,377
pattern recognition	0.22	0.05	7,286	22	7,111
…	…	…	…	…	…

Table 5 Cluster vertices in *G*(CPCL)

Vertex	Centrl.	Dens.	Betw.	Degre	w_Betw
machine learning	0.035	0.003	7,224	35	7,845
text mining	0.055	0.005	4,854	26	3,867
pattern recognition	0.196	0.064	3,141	18	3,069
classification	0.059	0.002	2,902	17	1,318
bioinformatic	0.083	0.020	2,607	16	1,370
…	…	…	…	…	…

The graphs *G*(*A*>*s*) and *G*(CPCL) constitute data analysis levels. For example, using the AiSee interactive interface (http://www.aisee.com), we can visualized the graph of clusters *G*(CPCL) that reveal the keywords that have the highest score of betweenness centrality since they are used as cluster labels. Opening the clusters, we access to the main pair of non central concepts related by geodesic paths that cross the label of the clusters suggesting potential new interactions between concepts. The higher betweenness and degree values (see tables 4 and 5) allow characterizing hubs and crossroad vertices in the graph structure. In the Appendix, by way of example two excerpts of *G*(CPCL) are presented where clusters within circles -text mining, bioinformatics, proteomics, machine learning, and classification- are hubs and crossroad at the same time in the global structure of the clustered graph. In cluster *G*(*A*>*s* = 0.001) the keywords labelling the clusters also are hubs and crossroad items. In both graphs *G*(*A*>*s* = 0.001) and *G*(CPCL), the data and text mining domain appears for the period observed strongly related to bioinformatics.

6. Conclusion and Future Work

In this article we have proposed a graph clustering methodology for mining the topic structure of an unordered set of textual data, which also can be browsed. The aim is to discover useful knowledge from an unordered dataset, i.e. useful for watching a research domain. We believe that the hypergraph modelling and the introduction of minimal traversal in the data analysis constitute something new in the use of the graph theory in informetric studies. The studies which we know concentrate principally on citations and then working with directed graph models, and do not refer any to hypergraph modelling; in addition co-word analysis is omitted.

The advantages of the hypergraph theory widely are that it constitutes a uniform data model to generalise co-word analysis, allowing to cover the entire model called relational (Callon et al, 1993). On the other hand, we shall compare it to more symbolic methods as formal concept analysis (FCA) using the Galois lattice model (Ganter et al, 2005; Ganter & Wille, 1999). Note that the number of possible closed sets is exponential on the number of attributes, and the generation of the whole Galois lattice is a NP hard problem. We are interested in clustering methods that naturally highlight particular small closed sets of attributes in linear time. Our intention is to try to point out the theoretical requirements of clustering algorithm in linear or quadratic time to maximize the probability of extracting closed sets.

A tree can be viewed as a graph with no circle of length higher than 2, then a natural generalization of trees are graphs with no circle longer than 3. This class of graphs is called chordal. We experimentally observed that all subgraphs with less than 30 elements of a co-author graph are chordal, meanwhile the whole graph is not. On the contrary, the co-word graph is more general and does not seem to have special chordal properties. In future work we shall study the chordal properties in the purpose of improving the graph visualisation.

Graph decomposition: to know that a graph is SWG allows to consider a “graph decomposition process” (Berri et al, 2004) using group of vertices as “clique minimal separators” aiming to split the SWG into not disjoint minimal unites; we are working on this.

References

Berge C. (1987) *Hypergraphes*, Paris: Gauthier-Villars.

Berry A., Kaba B., Nadif M., SanJuan E., Sigayret A. (2004) Classification et désarticulation de graphes de termes in JADT 2004 Proceedings, Leuven, Belgium, 10-12 march, p. 160-170.

Callon M., Courtial J-P., Penan H. (1993). *La Scientométrie.* Paris, Presses Universitaires de France, (coll. Que sais-je? Vol. 2727).

Courtial J-P. (1990) *Introduction à la scientométrie*. Paris: Anthropos – Economica.

Ganter B., Stummed G., Wille R. (Eds) (2005) Formal Concept Analysis: Foundations and Applications. Lecture Notes in Artificial Intelligence 3626, Springer.

Ganter B., Wille, R., (1999) Formal Concept Analysis: Mathematical Foundations, Springer.

SanJuan E., Dowdall J., Ibekwe-SanJuan F., Rinaldi F. (2005) A symbolic approach to automatic multiword term structuring, *Computer Speech and Language*, vol 19, 4, October 2005, p. 524-542.

Van Cutsem B. (Ed.) (1994) *Classification and Dissimilarity Analysis.* (LNS 93). Berlin: Springer.

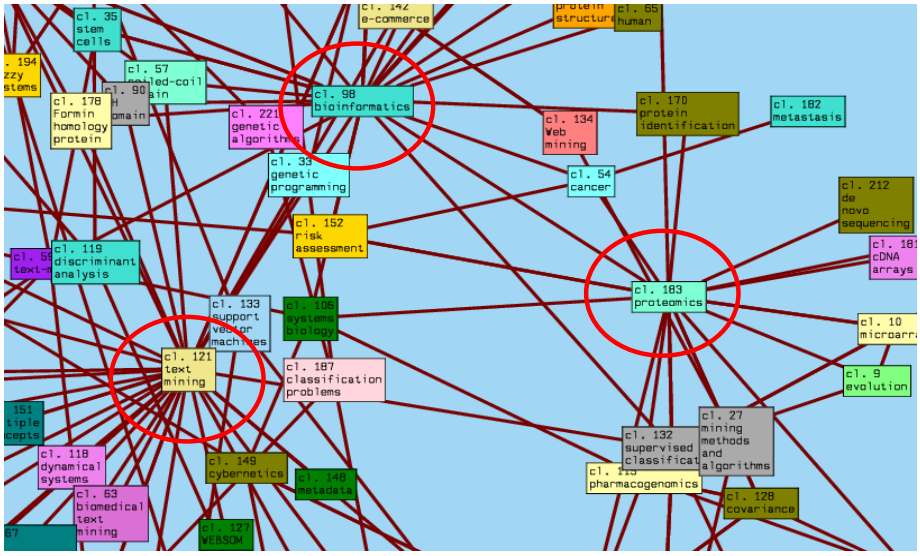
Watts D.J. (1999). Small Worlds, Princeton University Press

Wille R. (1982) Restructuring lattice theory: an approach based on hierarchies of concepts, in I. Rival (Ed.), *Ordered Sets*, vol. 83, D. Reidel, Dortrecht, p. 445-470.

Zaki M. (2004) Mining non-redundant association rules, Data Mining and Knowledge Discovery, 9 (3) p. 223-248

APPENDIX

(A) Text Mining, Bioinformatics, Proteomics



(B) Machine learning, Classification