Measuring group cohesion in document collections

Benjamin Renoust, Guy Melançon, Marie-Luce Viaud

Measuring group cohesion in document collections

Benjamin Renoust*†, Guy Melançon*
*CNRS UMR 5800 LaBRI, INRIA Bordeaux Sud-Ouest
Campus Université Bordeaux I
Talence, France
{Benjamin.Renoust, Guy.Melancon}@labri.fr

Marie-Luce Viaud†
†Institut National de l'Audiovisuel (INA)
Paris, France
mlviaud@ina.fr

***Abstract*—Exploring document collections remains a focus of research. This task can be tackled using various techniques, typically ranking documents according to a relevance index or grouping documents based on various clustering algorithms. The task complexity produces results of varying quality that inevitably carry noise. Users must be careful when interpreting document relevance or groupings. We address this problem by computing cohesion measures for a group of documents confirming/infirming whether it can be trusted to form a semantically cohesive unit. The index is inspired from past work in social network analysis (SNA) and illustrates how document exploration can benefit from SNA techniques.**

***Keywords*-document collection exploration, document group cohesion, entanglement index, text analytics**

I. Introduction

This paper adopts a social network analysis perspective to investigate the notion of group cohesion in a document collection. Our contribution exploits structural features of a network describing how index terms associated with documents interact with one another. Different indices are computed to measure *how well* and *how much* terms co-occur with one another. This perspective differs from and complements more traditional approaches such as document ranking and topic detection.

Most search engines rank the documents they deliver when queried using keywords. Document ranking techniques, among which Pagerank [4] is the likely most well known and widely used, most often order according to their *relevance* (with respect to a query) [7]. Ranking, however, does not necessarily consider relationships between keywords and terms indexing documents, leading to an ordered list of documents that concern distinct and often uncorrelated topics.

Another line of research focuses on the identification of *topics* in a document collection. Many approaches rely on terms to index documents and compute statistics indicating how relevant and important a word or topic is in a document or collection [18], [12], [21], [2], [3], [11].

The contribution of this work complements ranking techniques and topic identification approaches. Inspired by social network analysis [5], we introduce a *term interaction network* to measure how much terms indexing documents interact with one another within a document group. We refer to term interaction as *entanglement*, and compute entanglement *intensity* and *homogeneity*. These two statistics can then be used to assess the overall cohesion among a group of documents.

The term interaction network can be formed from any set of relevant terms that have been identified within a document group using any method (e.g., statistics [18], [2], LSA [12], [11] or LDA [3]). One particularly interesting use of the interaction network and entanglement measures is to provide feedback on any document grouping or clustering, allowing users to locate more cohesive subgroups that result from an automated grouping procedure or algorithm.

The next section briefly reviews related work before laying out the necessary definitions and notations in Section II to introduce the term interaction network and entanglement measures. A case study is then discussed to show the potential use of these devices (Section III). We demonstrate the interaction network and entanglement measures on data obtained from TV news excerpt manually indexed by documentalists at INA[1].

A. Related work

Document collection analysis classically considers a co-occurrence matrix, from which several indices can be derived. A well-known index is the tf-idf index [18], which computes a weight for terms. Documents d, d' can then be seen as vectors of weights indexed by terms, corresponding to a line in the co-occurrence matrix. These vectors can be used to evaluate similarities or dissimilarities between documents. The *cosine similarity* $\cos(d, d') = \frac{\langle d, d' \rangle}{\|d\| \|d'\|}$ is one well known and vastly used similarity index.

Since the seminal work of Salton [18], researchers have proposed improvements to the "bag-of-words" model (see [16], for instance). Latent Semantic Analysis (LSA) [12], or Latent Semantic Indexing [8], exploits the idea that words with similar meaning occur close together in text. These methods evaluate semantic proximity by performing singular

[1]INA is the French National Multimedia Institute, see `www.ina.fr`

value decompositions on a word count matrix. Probabilistic Latent Semantic Indexing (PLSI) [21] is based on a mixture decomposition derived from a latent class model that can be adjusted using an expectation-maximisation algorithm. Latent Dirichlet Allocation (LDA) [3] is a topic model similar to PLSI, where each document is viewed as a mixture of various topics. LDA assumes that each document is a mixture of a small number of topics, where the presence of words in documents is attributable to one of the document's topics.

These models share the common goal of finding the most relevant terms or topics emerging from a set of documents. Documents can then be described using either a weighted vector or probability distribution indexed by terms, thus allowing the user to compute similarities between documents. These weighted vectors can then feed different algorithms to mine and/or cluster document collections ([14], [24] or [20]).

Other authors have developed approaches focusing on *lexical cohesion* in relation with document ranking [22] or text structure [17]. In [22], the authors improve document ranking by computing various distances between terms. They exploit the structure of sentences and collocation of query terms in documents. Our work does not look at document content but assumes documents have already been indexed, or terms have already been extracted from documents. In our case, collocation of terms occurs when two terms index a same document. We use a term interaction network to capture multi-partite collocation, instead of considering only pairwise distance between terms.

Our approach also differs from these indexing and ranking techniques in various ways. First, we consider the term interaction network as a central ingredient from which the entanglement index is derived and several conclusions can be drawn. Our approach is similar to [1] because it considers a term-document network rather than the topic-term matrix used in LDA. The authors in [1] used a document-topic matrix to estimate the actual number of topics present in a document collection. Our concern is different, as we aim to establish whether a document group indeed forms a cohesive group for a given set of index terms. The network shape, however, may be a good indicator of the actual number of different topics that mix within a document collection.

The entanglement index may be computed on *any* group of documents and *any* term set indexing these documents. Our technique thus appears as a post-process, providing feedback about any indexing and/or grouping procedure used on a set of documents. The entanglement index is based on interactions that occur between terms (Section II) and fully exploits the interaction network topology. Our work shows how information retrieval can benefit from ideas and techniques borrowed from social network analysis (SNA). To our knowledge, most papers taking advantage of SNA in information retrieval do so by considering a social network of (human) actors, as in [13].Our work takes a completely different perspective and directly applies SNA to a network of terms seen as interacting entities.

II. Document group cohesion

We now define the entanglement index of terms based on matrix analysis of a term interaction network. Let D be a collection of documents $d \in D$, each indexed by terms $t \in T$, where T denotes a collection of terms. *Terms* here *index* documents and correspond to words either taken from a fixed vocabulary (thesaurus) or extracted from documents. An example is a video document (e.g., an excerpt from an evening TV news program) indexed by terms related to the news excerpt topics. We assume here that terms have already been identified and/or computed, so all documents come equipped with a set of index terms. Let $M = (m_{d,t})_{d \in D, t \in T}$ denote the usual co-occurrence matrix, where $m_{d,t}$ denotes the number of occurrences of term t in document d. Document d can then be seen as a vector of weights indexed by terms $t \in T$, namely $d = (m_{d,t})_{t \in T}$, corresponding to a line in the co-occurrence matrix M.

A. Term interaction network

One can define a graph-based representation of the document-term relations. The co-occurrence matrix indeed corresponds to a graph $G_{D,T} = (V, E)$, whose vertices are either documents or terms, $V = D \cup T$ and edges $e = \{d, t\} \in E$ connect documents to terms whenever $m_{d,t} > 0$. This graph is obviously *bipartite*, as edges never directly connect any two documents or any two terms. The top image in Fig. 1 illustrates this construction from a set of four different documents with index terms; the next image corresponds to the bipartite graph defined from these documents and index terms.

Many techniques and algorithms are found in the literature for exploiting this bipartite graph to mine and cluster either the document collection [23] [6], term set [19] or both simultaneously [9].

This bipartite graph is sometimes used to derive a graph $G_D = (D, E_D)$, directly linking documents. The graph is built from $G_{D,T}$ by projecting paths $d - t - d'$ onto edges $e = \{d, d'\} \in E_D$. Fig. 1 (third image from top) illustrates how G_D is obtained from $G_{D,T}$ (it *does not include loops* connecting a document to itself). Distinct terms $t, t', \ldots$ may link documents d and d' in $G_{D,T}$, from which the edge $e = \{d, d'\}$ is induced. We collect all such terms and turn them into attributes of edge e. Terms t may also be seen as types (labels) for edges in E_D. These terms are called the *terms associated with* the edge $e \in E_D$, common to both documents d and d'. Other applications turn this term set into a weight on edges $e \in E$ and exploit it when mining the document-document graph $G_D = (D, E_D)$. However, projecting a bipartite graph onto a one-mode network unfortunately leads to information loss

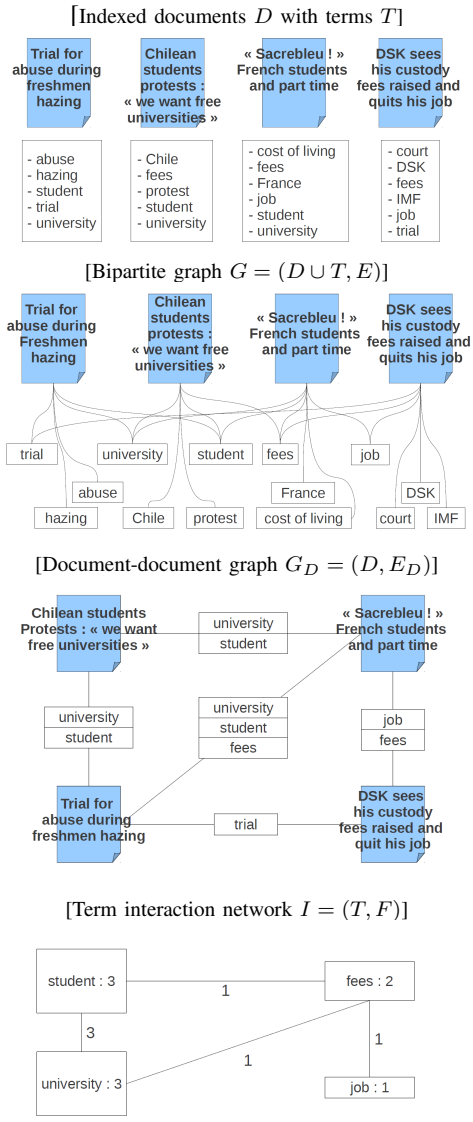

Figure 1. Graphs are built from a document collection indexed by terms (top), from which a bipartite graph linking documents and terms can be considered (next image). We then consider the projected document-document graph with terms as edge attributes (third image from top) and derive the resulting *term interaction network* (bottom).

[15], [25]. We overcome the situation by considering an additional network on terms.

We now build a *term interaction network*. Consider the graph $I = (T, F)$, whose vertices are terms $t \in T$. An edge $f \in F$ links terms $t, t' \in T$ whenever *they both are associated* with edge $e = \{d, d'\} \in E_D$, i.e., when two distinct documents $d, d' \in D$ are both indexed by terms t, t'. Terms t and t' *interact with one another* through at least two distinct documents d, d'.

The graph $I = (T, F)$ *is not* built from $G_{D,T}$ by projecting paths $t - d - t'$ onto edges $e = \{t, t'\}$. Fig. 1 (bottom) illustrates how I_T is obtained from graph G_D. The *term interaction network* $I = (T, F)$ is at the centre of our discussion and is actually an object of interest when exploring the document space. The interaction network is defined *after* documents have been indexed. The interaction network $I = (T, F)$ relies on the definition of a document-document graph $G_D = (D, E_D)$ that may be obtained from any data linking documents to terms. Nothing prohibits customizing the graph G_D before it is used to define the interaction network $I = (T, F)$.

The idea of building the term interaction network is borrowed from social network analysis [5]. We compute *interaction matrices* $N_I = (n_{t,t'})$ and $C_I = (c_{t,t'})$ (where subscripts are terms $t, t' \in T$). Let $e \in E_D$ be an edge in G_D and $\tau(e) \subset T$ denote the set of terms associated with e. Conversely, let $\tau^{-1}(t)$ denote the set of edges $e \in E_D$ with term $t \in T$ as an associated term. We write $n_{t,t} = |\tau^{-1}(t)|$ for the cardinality of that set. We also define $n_{t,t'}$ as the number of edges $e \in E_D$ with $\{t, t'\} \subset \tau(e)$, i.e., $n_{t,t'}$ equals the number of edges $e \in E_D$ that carry both terms t and t'. In other words, $n_{t,t'} = |\tau^{-1}(t) \cap \tau^{-1}(t')|, t \neq t'$.

Define $c_{t,t} = \frac{n_{t,t}}{|E_D|}$, and $c_{t,t'} = \frac{n_{t,t'}}{n_{t,t}}, t \neq t'$. If matrix N_I is symmetric, matrix C_I is not. The diagonal entries $c_{t,t}$ in matrix C_I can be informally seen as the probability that an edge in E_D is associated with term t. Non-diagonal entries $c_{t,t'}$ would then correspond to the conditional probabilities that an edge is associated with term t' given that it is associated with term t.

Consider the matrices N_I (left) and C_I (right) shown below. These matrices are computed from the 5-term clique in Fig. 2 indexing a collection of 18 documents sharing 103 links. Reading the diagonal, $n_{1,1} = 71$ links are associated with the first term (road safety), and $n_{2,2} = 48$ with the second (accident prevention). The number of links associated with both the first and second terms is $n_{1,2} = n_{2,1} = 35$. Reading the first entries in C_I, the first term is associated with $c_{1,1} = 69\%$ of all links, and there is a $c_{1,2} = 73\%$ chance of finding a link associated with the second term among those associated with the first term, while only $c_{2,1} = 49\%$ of the links are associated with the second term among those associated with the first term.

$$\begin{bmatrix} 71 & 35 & 61 & 46 & 28 \\ 35 & 48 & 35 & 41 & 15 \\ 61 & 35 & 78 & 42 & 45 \\ 46 & 41 & 42 & 67 & 21 \\ 28 & 15 & 45 & 21 & 45 \end{bmatrix} \quad \begin{bmatrix} 0.69 & 0.73 & 0.78 & 0.69 & 0.62 \\ 0.49 & 0.47 & 0.45 & 0.61 & 0.33 \\ 0.86 & 0.73 & 0.76 & 0.63 & 1 \\ 0.65 & 0.85 & 0.54 & 0.65 & 0.47 \\ 0.39 & 0.31 & 0.58 & 0.31 & 0.44 \end{bmatrix}$$

B. *Entanglement index*

We now wish to compute the *entanglement index* λ_t for each term t, measuring how much t is *entangled* with other terms in the network $I = (T, F)$, seen as a measure of its contribution to the cohesion of a document group. Let λ denote the maximum entanglement index among all terms and γ_t denote the fraction that computes the entanglement

for term t. The entanglement index for term t can then be computed as $\lambda_t = \gamma_t \cdot \lambda$.

Now, the entanglement index of a term is reinforced through interaction with other well entangled terms. Having a probabilistic interpretation of the matrix entries $c_{t,t'}$ in mind, we postulate the following equation which defines the values γ_t:

$$\gamma_t \cdot \lambda = \sum_{t' \in T} c_{t',t} \gamma_{t'} \tag{1}$$

The vector $\gamma = (\gamma_t)_{t \in T}$, collecting values for all terms t, thus forms a *right* eigenvector of the transposed matrix C'_I, as Eq. (1) gives rise to the matrix equation $\gamma \cdot \lambda = C'_I \cdot \gamma$. The maximum entanglement index thus equals the maximum eigenvalue of matrix C'_I.

The entanglement indices λ_t are of lesser interest; we are actually interested in the relative γ_t values. Furthermore, we shall see how the entanglement vector γ and eigenvalue λ can be used to define global measures to help understand how cohesion in a document group takes place.

C. Entanglement intensity and homogeneity

This section introduces *entanglement intensity* and *entanglement homogeneity* as global network measures for the term interaction network $I = (T, F)$. Its topology provides useful information about how terms contribute to the cohesion of a document group.

The archetype of an *optimally cohesive document group* is when all documents are indexed by the exact same terms. Indeed, assume either that experts have manually indexed documents or that terms have been obtained through some automated procedure(s).

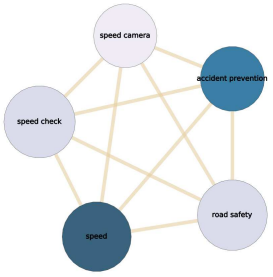

Figure 2. Optimal cohesion is reached when the entanglement of terms forms into a clique where all terms interact with one another with equal frequency.

The graph $I = (T, F)$ then corresponds to a *clique*, i.e., a graph where all nodes connect to all other nodes. In this case, all matrix entries $n_{t,t'}$ coincide, so all entries in matrix C_I equal 1. The maximum eigenvalue of C'_I then equals $\lambda = |T|$, and all γ_t are equal. That is, all terms indeed contribute, and they all contribute equally to the overall document group cohesion. The Perron-Frobenius theory of nonnegative matrices [10, Chap. 2] further shows that $\lambda = |T|$ is the maximum possible value for an eigenvalue of a non-negative matrix with entries in $[0, 1]$.

An opposite situation occurs when the term interaction network is not connected. Terms split into two subsets that never interact and suggests that the document set may be divided into distinct subgroups. Note that this type of information is easily revealed by inspecting the interaction network, although it is not immediately revealed when looking at a weight vector computed by most indexing techniques. In this case, the connected components in $I = (T, F)$ must be inspected independently. In the sequel, we assume $I = (T, F)$ is connected. In that case, non-negative matrix theory tells us that the matrix C_I has a maximal positive eigenvalue $\lambda \in R$ with a single associated eigenvector γ with non-negative real entries [10, Theorem 2.6].

Another typical situation occurs when few terms appear central and the remaining terms are peripheral. On the one hand, documents share a few common terms or rally around a few central topics; on the other hand, documents form subgroups around secondary terms or subtopics. This situation is again easy to identify, as the term interaction network has a fragmented structure (Fig. 4). In this case, the entanglement index reaches higher values for central terms while showing a clear decrease in peripheral terms. Peripheral subgroups may form smaller but denser subnetworks. When examining them locally and recomputing the entanglement index based on the term subset involved, we may expect the entanglement index to adjust to the clique scenario. The case studies presented below develop this situation.

Inspired from the clique archetype of an optimally cohesive document group, we wish to compute a entanglement index at the document group level. Because the eigenvalue is bounded above by $|T|$, we define the network *entanglement intensity* (or *intensity* for short) as the ratio $\frac{\lambda}{|T|} \in [0, 1]$ to measure how intensely interactions occur within a document group.

We also know that the clique situation with equal $c_{t,t'}$ weights leads to an eigenvector γ with identical entries. This eigenvector thus spans the diagonal space generated by the diagonal vector $1_T = (1, 1, \ldots, 1)$. This motivates the definition of a second measure providing information about how homogeneously interaction occurs among terms. We may indeed compute the cosine similarity $\frac{\langle 1_T, \gamma \rangle}{\|1_T\| \|\gamma\|} \in [0, 1]$ to get an idea of how close the document group is to being optimally cohesive. We will refer to this value as the network *entanglement homogeneity* (or *homogeneity* for short).

III. Use Cases

This section discusses two use cases illustrating how the various entanglement measures we have defined, together with the term interaction network topology, can be employed to explore a document collection and assess the cohesion of document groups.

Both use cases are built using TV news excerpts that cover many subjects over a 100-day period. Documents were manually indexed at INA. Document groups were identified using classical clustering approaches, outputting groups of varying sizes and homogeneity. The procedures used to form these groups are not the focus here. Indeed, the entanglement index and interaction network are designed to provide feedback about the groups returned by *any* grouping procedures. A term interaction network may be inferred from any document group one wishes to inspect.

A. Road Safety

We first consider a set of approximately 20 documents, all relating to road safety. Although small, this document sample exhibits interesting features that can also be found in larger samples. Road safety became a topic of interest after the government established a safety policy promoting the use of automated radar, with an inevitable increase in traffic tickets and fines. As expected, this news generated attention, and all TV channels devoted parts of their news programs to this subject. Documents involve index terms, including *accident prevention*, *arrest*, *danger*, *driver*, *driving behavior*, *money*, *offense*, *policeman*, *prison*,*road safety*, *society*, and *speed*, *speed camera* as the resulting interaction network shows in Fig. 3.

Darker nodes have higher entanglement indices. Node size relates to the number of links $n_{t,t}$ in G_D that are associated with term t. As we may guess from the layout, central nodes *accident prevention* and *speed* have higher entanglement indices, 0.38 and 0.44, respectively. Other nodes have lower values, such as *danger* and *speed camera* with 0.30 and 0.07, respectively. The intensity for the whole network $I = (T, F)$ is $\lambda/|F| = 0.33$, while the homogeneity is 0.81.

Fig. 3 clearly shows that terms split into subgroups indicating why optimal cohesion might not be reached. The interaction matrix N_I accordingly has a block structure (greyed background), with corresponding null off-diagonal blocks. The central terms are centred in the matrix and appear on top of a blue background. The matrix values show how terms interact within their components, except for central terms, which interact with all other terms. The upper part of the matrix corresponds to the upper part of the network and clearly shows that all terms interact with one another with low frequencies (e.g., terms index a small subset of all documents). The lower part of the matrix exhibits a completely different behaviour, where terms interact more vividly, but not with all other terms in the component.

The network topology suggests closely examining documents relating to terms at the bottom, positioned below the central terms *accident prevention* and *speed*. We consider a subgraph I' formed with the terms *fine*, *policeman*, *speed camera*, and *road safety*, etc. Terms in I' index all documents but one. For this sub-network I', we have an

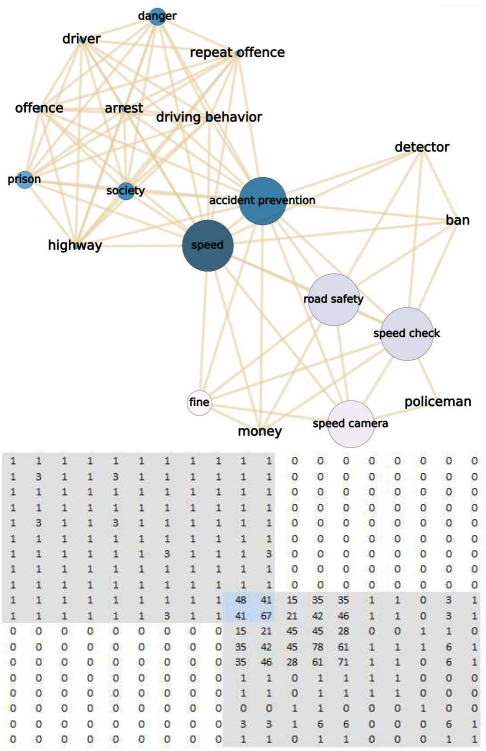

Figure 3. Interaction network induced from a document set related to road safety and speed cameras. The network splits into two overlapping components organised around two central terms: *road safety* and *speed*, leading to an obvious block structure for the interaction matrix.

entanglement intensity of 0.31 and homogeneity of 0.72, which is below those of the total network I. These lower values occur because many terms, e.g., *fine*, *money* and *detector*, actually index few documents (as suggested by their sizes) and the terms involved in I' distribute less evenly among documents than all terms in I globally do, as revealed by the zero entries in the lower right part of the matrix.

We may further discard the low interaction terms and focus on the 5-node clique in Fig. 2 associated with 18 of the 20 documents. The clique corresponds to the submatrix with high integer values near the centre of the image (Fig. 3). As expected, the clique reaches higher intensity 0.6 and homogeneity 0.98, confirming that these 18 documents form a cohesive unit around the five selected terms.

This first example clearly shows that entanglement indices, intensity and homogeneity provides insight on the structure of the document group. This document group clearly is homogeneous as all documents after all concern road safety. Different facets of the group may be investigated using techniques such as LDA to identify *topics* [3], with additional efforts to evaluate the actual number of topics [1]. However, the shape of the network topology is a clear indicator of the number of "topics" or "stories" the group contains, as confirmed when observing the increased homogeneity of these two components.

Also, LDA provides no information on the relational structure underlying the term set.

Local inspection of the term interaction network moreover points at marginal terms that may be discarded in order to find core terms around which cohesion builds.

B. Students

The second example concerns a group of 36 documents about students and universities. They gather 3 stories about student protests in Chile, excessive behaviour during freshmen hazing and students' financial conditions, among other diverse related subjects. Documents are indexed using many terms, including *Chile*, *Grande Ecole*, *higher education*, *cost of living*, *education*, *hazing*, *protests*, *salary*, *student*, *university*, and *violence*, etc. Fig. 4 shows the resulting interaction network.

The terms with highest entanglement indices sit in the centre of the network: *student* and *higher education*, with respective values 0.72 and 0.47. The terms *hazing* and *university* immediately follow with entanglement indices of 0.34 and 0.27, respectively. Compared to the the previous example, the network globally has a low intensity of 0.09 and homogeneity 0.44. This is due to a more intricate topology splitting into three denser regions together with satellite terms.

Let us now focus on denser areas in the quest for more cohesive subgroups. The lower part of the network organises around the term *protests* and links to documents related to student protests in Chile and London. These terms (with 10 related documents) lead to a sub-network with increased intensity 0.23 and homogeneity 0.81. Further focusing on *Chile* leads to a 5-node clique with slightly higher intensity 0.30 and homogeneity 0.90. After foreign newspapers reported student protests in Chile and London, the French press raised interest about the condition of French students in universities. The terms *salary*, *cost of living* and *part time* thus link within the network (rightmost area of the network in Fig. 4), as they index documents concerned with students' life conditions both in France and abroad. Focusing solely on these three terms, we get a smaller sub-network (and four associated documents) with much higher intensity 0.64 and homogeneity 0.93; again, smaller term and document subsets typically reach higher cohesion. The upper part of the network in Fig. 4 gathers documents related to initiation rites in *Grande Ecole*. The press got interested in these rites after freshmen students complained about abuse during the hazing rites and brought their cases to court. The four terms at the top (with related documents) induce a sub-network with intensity 0.41 and homogeneity 0.76. Although intensity and homogeneity are higher than the overall network, they remain far from the optimal case and can be considered low. When more closely examining the situation, these low values occur because terms are *nested*: edges in G_D associated with *law* are all associated with *Grande Ecole*, which are all associated with *customs*, ultimately contained in the set of edges associated with *hazing*.

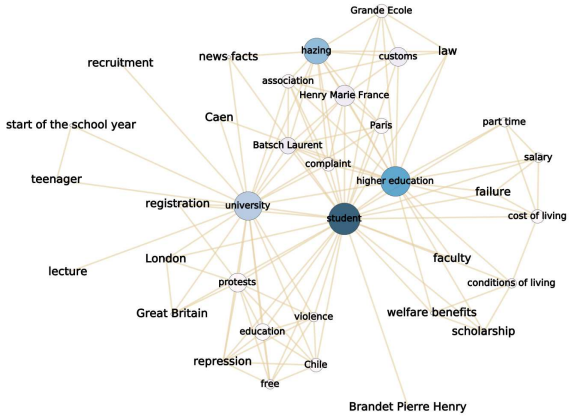

Figure 4. Interaction network induced from a set of documents related to students and universities. This network has a much more intricate topology and twice as many terms as the previous example (Fig. 3).

This perfectly exemplifies how network topology can guide the document collection exploration. Identifying denser areas in the network is a useful strategy for selecting more cohesive documents from within the originally queried document set.

C. Structure of the (Intensity × Homogeneity) space

Now that we have put forward how the interaction network supports storytelling, and entanglement indices support term ranking, we can conclude this section and return to entanglement profiles considering the previously discussed examples. We have used intensity $\frac{\lambda}{|F|}$ and homogeneity $\frac{\langle 1_T,\gamma\rangle}{\|1_T\|\|\gamma\|}$ as two distinct measures to provide complementary information about the term interaction network. The examples exhibit situations where intensity and homogeneity can be either low or high. Although we may suspect that these quantities do not vary independently, we nevertheless design a 2D plot where homogeneity is plotted along the x-axis and intensity is plotted along the y-axis (Fig. 5). Any term interaction network would then be plotted as a 2D point $(x, y) = (\frac{\langle 1_T,\gamma\rangle}{\|1_T\|\|\gamma\|}, \frac{\lambda}{|F|})$ in the plane, and we may expect the plot to divide into areas that correspond to network profiles.

There is an obvious dependency between intensity and homogeneity: high intensity cannot be achieved without some amount of homogeneity so the space cannot spread

The clique was presented as the archetype of an optimal interaction network located at the top rightmost position $(x, y) = (1, 1)$ in the plot. The top-right area thus collects these relatively dense and evenly interacting networks. The 5-node clique in Fig. 2 falls into this profile category, as does the "Road safety" upper sub-network considered in the first use case (Fig. 3, Section III-A). Higher intensity

and homogeneity are much easier to achieve with smaller document and term subsets.

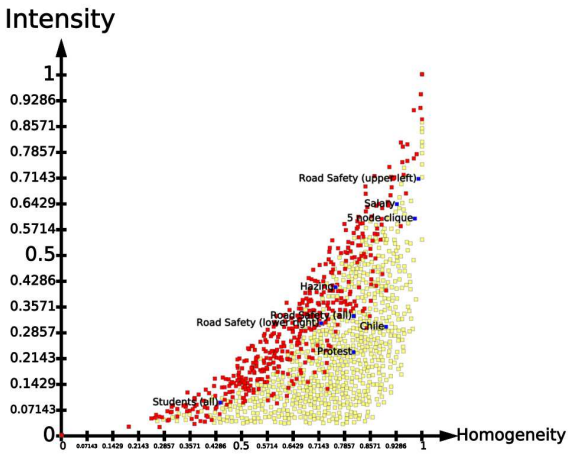

Figure 5. Entanglement profiles can roughly be categorised by combining intensity and homogeneity. Measures on INA's document groups are in red. Randomly generated bipartite graphs are in yellow. Use cases examples are in blue with labels.

The lowest-right area corresponds to relatively high homogeneity and lower intensity: terms almost all interact with one another, but not as much as the document graph G_D theoretically allows. N_I matrices are non-sparse, and they have large diagonal but rather low off-diagonal entries. Example networks in this category are the "Road safety" lower sub-network (Fig. 3, Section III-A), and the *Chile* sub-network with intensity 0.30 and homogeneity 0.90 is also part of this category.

The most left area is tricky. This case occurs when terms are nested, as if they were expressing similar concepts at different generality levels. This situation translates into consecutive inclusion of terms among documents links (i.e., links in G_D associated with term t include all links associated with term t' plus some other links). We pointed out the "Hazing" sub-network in the second use case as a prototype of this phenomenon. The fact that it nevertheless has intensity 0.44 and homogeneity 0.76 stresses the fact that the areas defined by the orthogonal dotted lines must be refined and/or revised.

The lower-left case gathers networks with low intensity and low homogeneity. This is a rather common case, usually gathering more documents and terms with loose interaction. This is a situation where many terms could appear as satellites of more central terms. A term set covering a wider semantic scope inevitably induces a network falling in this category. A typical network would have a low density (few edges) and a random link structure, leading to a sparse N_I matrix with ϵ entries. We could argue that the starting interaction network in the *Students* use case, with a intensity of 0.09 and a homogeneity of 0.44, falls within this category.

Although we have generated some random graphs to illustrate the span of homogeneity and intensity, more experimentation is needed to assess these prospective categories, determine thresholds defining the profile areas and estimate how they are populated.

IV. Conclusion

This paper introduced a term interaction network (Section II) as a device from which term entanglement indices can be computed. The cohesion indices can then be translated into global cohesion intensity and homogeneity measures among terms *in a group of documents*. The cohesion index, cohesion intensity and homogeneity can be computed for *any* group of documents. They can be used to provide feedback about procedures used to group documents, helping users decide whether documents can be trusted to form a genuine cohesive semantic unit.

The case study (Section III) clearly shows the added value brought by the interaction network topology. The network shape is a clear indicator of possible profiles with an obvious archetype profile of an optimally cohesive group as a clique. The examples show how the topology organises into different areas: some terms are deeply nested into a region, while others act as pivot between regions. Diagram 5 was used to distinguish four generic profiles induced from different cohesion intensity and cohesion homogeneity pairs $(\frac{\lambda}{|F|}, \frac{\langle 1_T, \gamma \rangle}{\|1_T\| \|\gamma\|})$.

Our technique is independent of the procedure used to extract or define the terms used to index documents; it thus usefully complements existing indexing techniques. LDA assumes that each document contains a mixture of topics that are revealed in a document collection as a mixture of terms. Determining the exact number of topics combined in a document collection is a difficult problem [1]. The case studies suggest that this number may correlate or be derived from the term interaction network shape. Denser sub-networks coupled with relatively high interaction weights $c_{t,t'}$ correspond to higher local cohesion. Although a document group may be loosely cohesive, the interaction network may lead to discovering more cohesive term and document subsets.

The examples used have relatively small sizes. The largest document samples we considered gather hundreds of documents and terms, but this limitation is apparent, as using the interaction network occurs after documents have been indexed and grouped. Although a query might return thousands of documents, we may expect the grouping procedure to form much smaller groups before closer examination occurs. We also suspect that larger document samples gather larger term sets, typically leading to sparser term interaction matrices. This is confirmed by the examples discussed in Section III. Conversely, a close examination of the term interaction network helps to identify the core terms from which documents form a cohesive unit.

We plan to examine strategies to automatically identify term and document subsets with optimal (maximal) cohesion intensity and homogeneity. These problems, however, will inevitably bring us to combinatorial optimisation problems, and we may expect to have no choice but to rely on heuristics to avoid typical algorithmic complexity issues.

REFERENCES

[1] R. Arun, V. Suresh, C. Veni Madhavan, M. Narasimha Murthy, M. Zaki, J. Yu, B. Ravindran, and V. Pudi. On finding the natural number of topics with latent dirichlet allocation: Some observations advances in knowledge discovery and data mining. volume 6118 of *Lecture Notes in Computer Science*, pages 391–402. Springer, 2010.

[2] R. Beaza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, New York, 1999.

[3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(5):993–1022, 2003.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[5] R. Burt and T. Scott. Relation content in multiple networks. *Social Science Research*, 14:287–308, 1985.

[6] D. Cai, X. He, and J. Han. Document clustering using locality preserving indexing. *Knowledge and Data Engineering, IEEE Transactions on*, 17(12):1624–1637, 2005.

[7] B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Addison Wesley, 2009.

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

[9] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 269–274, San Francisco, California, 2001. ACM.

[10] J. Ding and A. Zhou. *Nonnegative Matrices, Positive Operators and Applications*. World Scientific, Singapore, 2009.

[11] S. T. Dumais. Latent semantic analysis. In *Annual Review of Information Science and Technology (ARIST)*, volume 38, pages 189–230. 2004.

[12] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. Using latent semantic analysis to improve access to textual information. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 281–285. ACM, 1988.

[13] L. Kirchhoff, K. Stanoevska-Slabeva, T. Nicolai, M. Fleck, and K. Stanoevska. Using social network analysis to enhance information retrieval systems. *Applications of Social Network Analysis (ASNA), Zurich*, 7:1–21, 2008.

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

[15] M. Latapy, C. Magnien, and N. D. Vecchio. Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1):31–48, 2008.

[16] G. Lebanon, Y. Mao, and J. V. Dillon. The locally weighted bag of words framework for document representation. *Journal of Machine Learning Research*, 8(10):2405–2441, 2007.

[17] J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Comput. Linguist.*, 17(1):21–48, 1991.

[18] G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1983.

[19] N. Slonim and N. Tishby. Document clustering using word clusters via the information bottleneck method. In *23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 208–215. ACM, 2000.

[20] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *In KDD Workshop on Text Mining*, 2000.

[21] H. Thomas. Probabilistic latent semantic indexing. In *22nd ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, Berkeley, California, United States, 1999. ACM.

[22] O. Vechtomova and M. Karamuftuoglu. Lexical cohesion and term proximity in document ranking. *Information Processing & Management*, 44(4):1485–1502, 2008.

[23] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '03, pages 267–273. ACM, 2003.

[24] Y. Zhao, G. Karypis, and U. Fayyad. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10:141–168, 2005.

[25] T. Zhou, J. Ren, M. Medo, and Y.-C. Zhang. Bipartite network projection and personal recommendation. *Physical Review E*, 76(4):046115, 2007.