Mesurer la cohésion sémantique dans les corpus de documents

Guy Melançon, Benjamin Renoust, Marie-Luce Viaud

RESEARCH REPORT
N° 8075
September 2012
Project-Teams GRAVITE

ISSN 0249-6399 ISRN INRIA/RR--8075--FR+ENG

Mesurer la cohésion sémantique dans les corpus de documents

Guy Melançon, Benjamin Renoust, Marie-Luce Viaud

Équipes-Projets GRAVITE

Rapport de recherche n° 8075 — September 2012 — 18 pages

Résumé : L'exploration de corpus documentaire reste encore aujourd'hui un domaine actif de recherche. Cette tâche peut être abordé à l'aide de nombreuses techniques, s'appuyant typiquement sur le calcul d'indices de pertinence ou de regroupement thématique (clustering). Ces solutions sont souvent empreintes de bruit, du fait même de la complexité de la tâche à mener. Les utilisateurs se doivent par conséquent d'être précautionneux lorsqu'il s'agit d'interpréter les résultat d'un ordonnancement ou de regroupement thématique des documents. Nous nous penchons sur cette dernière question et calculons des indices de cohésion sémantique associés à un groupe de documents permettant de questionner la cohésion d'un groupe de documents. Ces indices s'inspirent de travaux passés en analyse des réseaux sociaux (SNA) et montre combien il semble possible d'exploiter les résultats de ce domaine à des fins d'exploration de bases documentaires.

Mots-clés : exploration de corpus documentaires, cohésion sémantique, analyse de textes

RESEARCH CENTRE
BORDEAUX – SUD-OUEST

351, Cours de la Libération
Bâtiment A 29
33405 Talence Cedex

Measuring semantic cohesion in document collections

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.

Key-words: document collection exploration, semantic cohesion, text analytics

1 Introduction

Handling multimedia document collections comprising text documents, images and/or videos is a task most users face in their daily work. Scientists daily browse bibliographic databases of papers, images or videos. Journalists query multimedia databases, searching for documents published about an event or covering a topic. Conversely, specialised media broadcasting corporations or institutes (e.g., the Associated Press, AFP, Reuters or major newspapers worldwide) offer searching and querying services in their document archives. Designing new techniques or complementing existing ones to support exploring and searching in a document corpus fulfils a real demand from users and industrial actors.

Most search engines rank the documents they deliver when queried using a keyword set. Document ranking techniques order documents based on a relevance notion (with respect to a query) [8], among which Pagerank [5] is the likely most well known and widely used. Ranking, however, does not directly consider semantics, leading to an ordered list of documents that concern distinct and not-necessarily correlated topics.

Identifying topics in a document collection is a central problem that has been addressed in many ways. 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 [17] [1]. Latent semantic analysis (LSA) [13] (see also [12]) computes vector-based text representations to capture a document's semantic content. Latent Dirichlet allocation (LDA) improves LSA [4] by defining an approach with a solid statistical foundation.

The contribution of this work complements these indexing techniques and topic identification approaches and introduces a weighted *term interaction network* as a useful device 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 (Section refsubsec :profile). This 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, LSA or LDA). Spectral analysis, inspired by social network analysis [6], is then conducted on the network to define an *entanglement index* assigned to each term. The overall distribution of these entanglement indices with the numerical properties of the term interaction matrix, is used to assess the cohesion of a group or subgroup of documents and terms. One particularly interesting use of the interaction network and entanglement index is to provide feedback on any document grouping or clustering, allowing users to locate more semantically cohesive subgroups that result from an automated grouping procedure or algorithm.

This paper briefly reviews related work before laying out the necessary definitions and notations in Section 2 to introduce the term interaction network and entanglement index. A case study is then discussed to show the potential use of these devices (Section 3). We demonstrate the interaction network and entanglement index on data obtained from a Multimedia Institute [1]. We also report a user study conducted with expert documentalists to assess the cohesion identified by the entanglement index. Section 5 finishes the paper with a discussion and conclusion.

1.1 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 [17], which computes a weight for terms. Documents d, d' can then be seen as a vectors of weights indexed by terms, corresponding to a line in the co-occurrence matrix. These vectors can be used to evaluate similarities or

[1] The institute and URL are kept hidden in this version to comply with the anonymity rule.

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 [17], researchers have proposed improvements to the "bag-of-words" model (see [3], for instance). Latent Semantic Analysis (LSA) [13], or Latent Semantic Indexing [9], exploits the idea that words with similar meaning occur close together in text. These methods evaluate semantic proximity by performing singular value decompositions on a word count matrix (i.e., document sections or paragraphs), thus producing a more accurate and reliable representation of a document as a weighted vector of terms. 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) [4] 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. Topic models allow probabilistic term frequency occurrence modelling in documents.

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 ([16], [23] or [20]). It is necessary here to distinguish between a topic and a term. In practice, LDA computes a *topic* distribution using a document collection, which is a probability distribution assigned to a set of words present in the document collection. We are only concerned with *terms*, which correspond to words found in documents or words borrowed from a controlled vocabulary used to index documents.

Our approach differs from these indexing techniques in various ways. First, we consider the term interaction network a central ingredient from which the entanglement index is derived and several conclusions can be drawn. Our approach is similar to [2] because it considers a term-document network rather than the stochastic topic-term matrix used in LDA. The authors in [2] 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 (Section 3).

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 2) 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 a study [15] from the ASNA Conference series [2] and papers from the SNA-KDD workshop [3]. A common trend is to run a document analysis technique coupled with knowledge extracted from a network of actors and exchanges. Our work takes a completely different perspective and directly applies SNA to a network of terms seen as interacting entities.

[2] www.asna.ch

[3] www.snakdd.com

2 Document group cohesion

We now turn to defining a entanglement index based on the spectral 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 topic. 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.

2.1 Term interaction network

One can define a graph-based representation of the document-term relations. The co-occurrence matrix indeed corresponds to 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. This graph is obviously *bipartite*, as edges never directly connect any two documents or any two terms. Figure 1 illustrates this construction from a set of four different documents with index terms (a). Figure 1 (b) corresponds to the bipartite graph defined from these documents and index terms.

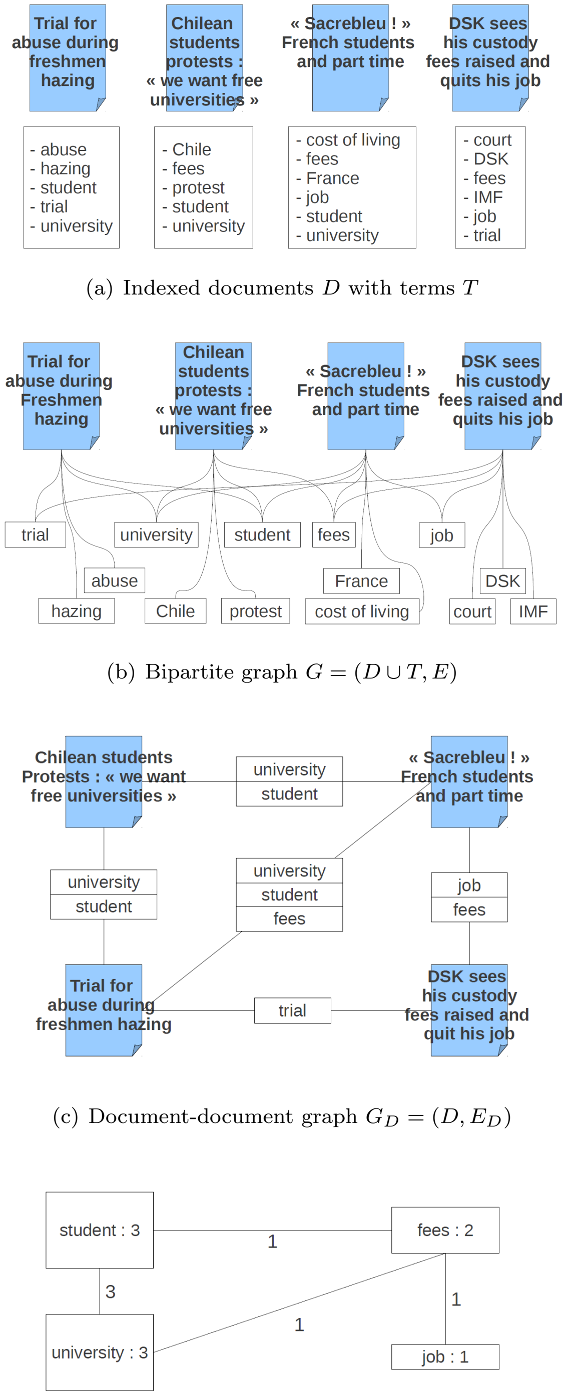
Edges $e \in E$ can be equipped with weights $\omega : E \to R$. An obvious candidate weight function is $\omega(e) = m_{d,t}$. Documents may also be indexed by terms using LDA, where $\omega(e) = P(t|d)$ could be a probability predicted by the LDA model. Any other weight function resulting from terms indexing documents may also be used. Many techniques and algorithms are found in the literature for exploiting this bipartite graph to mine and cluster either the document collection [22] [7], term set [19] or both simultaneously [10].

This bipartite graph is sometimes used to derive 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 $e = \{d, d'\} \in E_D$. Figure 1 (c) illustrates how G_D is obtained from $G_{D,T}$ (it *does not include loops* connecting a document to itself). Many distinct terms $t, t', \ldots$ may link documents d and d' in $G_{D,T}$, from which 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 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'. Many applications turn this term set into a weight on edges $e \in E$ and exploit it when mining the document-document graph $G = (D, E)$.

We now build a *term interaction network*. Consider graph $I = (T, F)$, whose vertices are terms $t \in T$. 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'.

Graph $I = (T, F)$ *is not* built from $G_{D,T}$ by projecting paths $t - d - t'$ onto edges $e = \{t, t'\}$. Figure 1 (d) 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 customising 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 [6]. We compute *interaction matrices* $N_I = (n_{t,t'})$ and $C_I = (c_{t,t'})$ (where subscripts are terms


Trial for abuse during freshmen hazing
Chilean students protests : « we want free universities »
« Sacrebleu ! » French students and part time
DSK sees his custody fees raised and quits his job
- abuse
- hazing
- student
- trial
- university
- Chile
- fees
- protest
- student
- university
- cost of living
- fees
- France
- job
- student
- university
- court
- DSK
- fees
- IMF
- job
- trial
Trial for abuse during Freshmen hazing
Chilean students protests : « we want free universities »
« Sacrebleu ! » French students and part time
DSK sees his custody fees raised and quits his job
trial
university
student
fees
job
abuse
France
DSK
hazing
Chile
protest
cost of living
court
IMF
Chilean students Protests : « we want free universities »
university
student
« Sacrebleu ! » French students and part time
university
student
university
student
fees
job
fees
Trial for abuse during freshmen hazing
trial
DSK sees his custody fees raised and quit his job
student : 3
1
fees : 2
3
1
1
university : 3
job : 1

(a) Indexed documents D with terms T

(b) Bipartite graph $G = (D \cup T, E)$

(c) Document-document graph $G_D = (D, E_D)$

(d) Term interaction network $I = (T, F)$

FIGURE 1 – Graphs are built from a document collection indexed by terms (a), from which a bipartite graph linking documents and terms can be considered (b). We then consider the projected document-document graph with terms as edge attributes (c) and derive the resulting *term interaction network* (d).

$t, t' \in T$). 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 $n_t = |\tau^{-1}(t)|$ for the cardinality of that set. We first 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, $n_{t,t'} = |\tau^{-1}(t) \cap \tau^{-1}(t')|$.

Define $c_{t,t} = \frac{n_{t,t}}{|F|}$, and $c_{t,t'} = \frac{n_{t,t'}}{n_t}$. 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 Figure 2, Section 2.3, 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 associated with the second term are also associated with the first term.

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

2.2 Entanglement index

We now wish to compute the *cohesion index* for each term, measuring how much a term t contributes to the overall cohesion of a document group. This notion of cohesion is directly adapted from a similar notion of social relation ambiguity [6]. Let λ denote the maximum cohesion index among all terms and γ_t denote the fraction that computes the cohesion for term t. The cohesion index for term t can then be computed as $\gamma_t \cdot \lambda$.

Because term cohesion is reinforced through interactions with other cohesive terms, having a probabilistic interpretation of the matrix entries $c_{t,t'}$ in mind, we can postulate the following equation which defines the values γ_t.

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

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 cohesion index thus equals the maximum eigenvalue of matrix C_I'.

The actual cohesion index values are of lesser interest ; we are actually interested in the relative γ_t values. Furthermore, we shall see how the cohesion vector γ and eigenvalue λ can be translated into a global measure to help understand cohesion in a document group. The next section introduces *entanglement intensity* and *entanglement homogeneity* as global network measures.

2.3 Cohesion profiles

The topology of the term interaction network $I = (T, F)$ provides useful information about how terms contribute to the overall semantic cohesion of document groups or subgroups. The focus here is on interactions among terms within the collection and aims to reveal how cohesive the collection is considering this set of terms.

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). Documents are thus identically indexed. This situation is one where either document experts or term extraction algorithms agree that all documents concern the same topics with the same "intensity", thus forming an optimally cohesive document group.

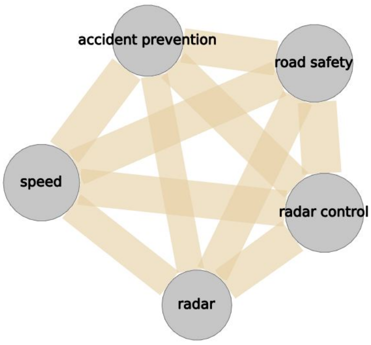


FIGURE 2 – An optimally cohesive document group turns the term interaction network into a clique where all terms interact with one another with equal intensity.

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 $\lambda = |F|$, and all γ_t coincide. That is, all terms indeed contribute, and they all contribute equally to the overall document group cohesion. The Perron-Frobenius theory of nonnegative matrices [11, Chap. 2] further shows that $\lambda = |F|$ 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. 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. This situation also indicates that the document set may be divided into subgroups, with the exception of a single document indexed using terms of distinct connected components.

The connected components in $I = (T, F)$ may thus be inspected independently. When $I = (T, F)$ is not connected, matrix C_I is considered reducible ; conversely, when $I = (T, F)$ is connected, the matrix C_I is *irreducible*. In that case, non-negative matrix theory tells us that matrix C_I has a maximal positive eigenvalue $\lambda \in R$ with a single associated eigenvector γ. This eigenvector has non-negative real entries [11, Theorem 2.6]. We hereafter assume that C_I is irreducible. 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 star-shaped structure (Figure 3). In this case, the cohesion index reaches higher values for central terms while showing a clear decrease in peripheral terms (nodes in Figure 3 are coloured according to their cohesion indices ; lower value nodes are lighter). Peripheral subgroups may form smaller but denser sub-networks. When examining them locally and recomputing the cohesion

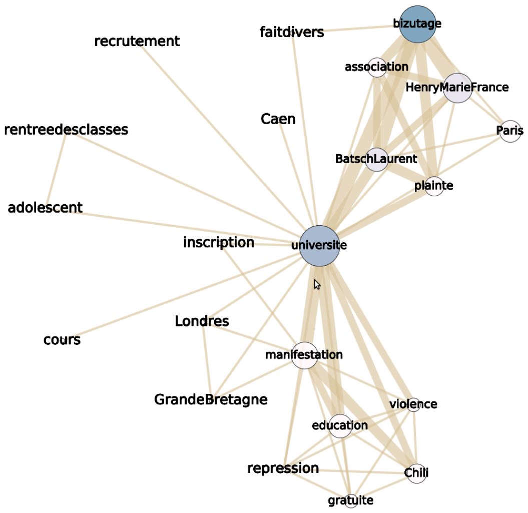


Figure 3 – A star-shaped interaction network gathers (a) central term(s), reaching a higher cohesion index. Peripheral terms may form denser subgroups with higher *local* cohesion.

index based on the term subset involved, we may expect the cohesion 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 cohesion index at the document group level. We already know that the eigenvalue is bounded above by $|F|$, so the ratio $\frac{\lambda}{|F|} \in [0, 1]$ measures how intense the interaction is within a document group. This ratio provides a measure for *entanglement intensity* among all documents *with respect to terms in* T.

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 cohesion distributes homogeneously 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 *cohesion homogeneity*. Other measures, including Shannon entropy [18] and Guimera's participation coefficient [14], offer interesting alternatives to cosine similarity.

3 Use Cases

This section discusses two use cases illustrating how cohesion measures and term interaction network topology can be employed to explore a document group.

Both use cases are built using TV news excerpts that cover many subjects over a 100-day period. Documents were manually indexed at the (anonymized) "Multimedia Institute", from which we could compute a co-occurrence matrix. 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 cohesion index and interaction network are designed to provide feedback about the groups returned by *any* grouping procedures. An interaction network may be inferred from any document group one wishes to inspect.

3.1 Road Safety and Radars

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*, *radar*, *road safety*, *society*, and *speed* (we list them here in alphabetical order). Figure 4 shows the resulting interaction network (the index terms are in French).

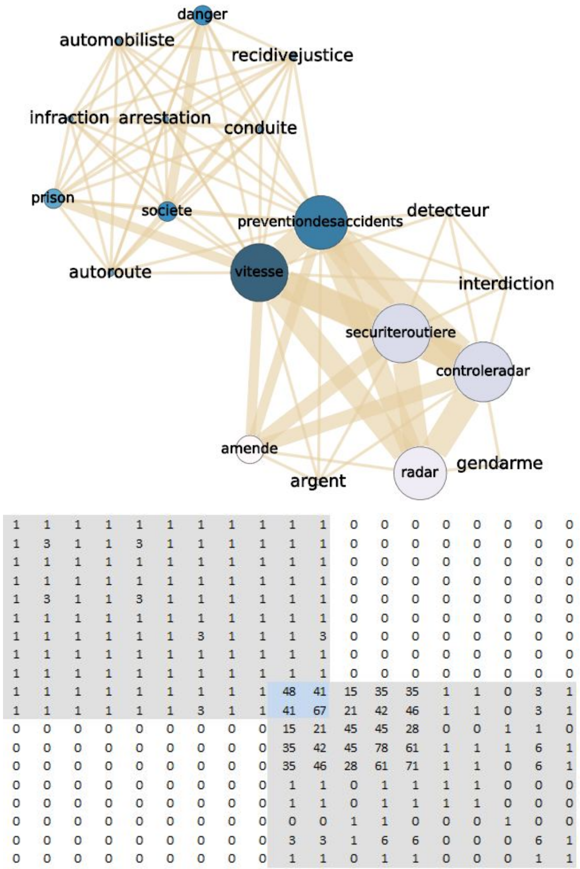
Darker nodes have higher cohesion indices. Node size relates to the number of links *in the graph* G_D that are associated with terms. As we may guess from the layout, central nodes *accident prevention* and *speed* have higher cohesion indices, 0.38 and 0.44, respectively. Other nodes have lower values, such as *danger* and *radar* with 0.30 and 0.07, respectively. The cohesion intensity for the whole network $I = (T, F)$ is $\lambda/|F| = 0.33$, while the cohesion homogeneity (cosine similarity, Section 2.3) is 0.81.

Figure 4 clearly shows that terms split into subgroups indicating why optimal semantic cohesion might not be reached. 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 the 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 completely different behaviour, where terms interact more vividly, but not with all other terms in the component.

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 *amende* (fine), *gendarme* (policeman), *radar*, and *sécurité routière* (road safety), etc. Terms in I' index all documents but one. For this sub-network I', we have a cohesion intensity of 0.31 and cohesion homogeneity of 0.72, which is below those of the total network I. These lower values occur because many terms, e.g., *amende*, *argent* (money) and *detecteur* (sensor), 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. (Note that the non-linearity of the $\cos(-)$ function must be considered when interpreting the increase in cosine similarity.)

We may further discard the low interaction terms and focus on the 5-node clique in Figure 2 (Section 2.3) associated with 18 of the 20 documents. (The clique corresponds to the submatrix with high integer values near the centre of the image.) As expected, the clique reaches higher cohesion intensity 0.6 and cohesion homogeneity 0.98, confirming that these 18 documents form a cohesive semantic unit around the five selected terms.

We close this first example by looking at the upper part of the network, formed with terms positioned above the central terms *accident prevention* and *speed*. We consider a subgraph formed with the terms *danger*, *automobiliste* (driver), *autoroute* (highway), and *prison*, etc. We get the sub-network I'' sitting at the top of Figure 4, where terms index 19 of the 20 original documents. By restricting the analysis to this sub-network I'', cohesion intensity and cohesion homogeneity are 0.57 and 0.98, respectively. After discarding the central nodes *accident prevention* and *speed*, cohesion intensity and cohesion homogeneity increase to 0.71 and 0.99, respectively. Again, we have a semantically cohesive unit, which incidentally forms a clique with uneven interaction



1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	3	1	1	3	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	3	1	1	3	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	1	1	1	1	1	3	1	1	1	3	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	48	41	15	35	35	1	1	0	3	1
1	1	1	1	1	1	3	1	1	41	67	21	42	46	1	1	0	3	1
0	0	0	0	0	0	0	0	0	15	21	45	45	28	0	0	1	1	0
0	0	0	0	0	0	0	0	0	35	42	45	78	61	1	1	1	6	1
0	0	0	0	0	0	0	0	0	35	46	28	61	71	1	1	0	6	1
0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0	0	0
0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	3	3	1	6	6	0	0	0	6	1
0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	1	1

Figure 4 – Interaction network induced from a document set related to road safety and radars. The network obviously splits into two components organised around two central terms : *road safety* (prévention des accidents) and *speed* (vitesse), leading to an obvious block structure for the interaction matrix (assuming terms are properly ordered).

weights $n_{t,t'}$. This conclusion must, however, be moderated because terms in this smaller sub-network index only four documents. Higher intensity and homogeneity are much easier to achieve with smaller document and term subsets.

3.2 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*, *Grandes Ecoles*, *higher education*, *cost of living*, *education*, *hazing*, *protests*, *salary*, *student*, *university*, and *violence*, etc. (listed here in alphabetical order). Figure 5 shows the resulting interaction network (terms are in French in the figure).

The terms with highest cohesion indices sit in the centre of the network : *students* (étudiant) and *higher education* (enseignement supérieur), with respective values 0.72 and 0.47. The terms *hazing* (bizutage) and *university* (université) immediately follow with cohesion indices of 0.34 and 0.27, respectively. The network has a much more intricate topology and twice as many terms as the previous example, which prohibits directly analysing its matrix. (The sub-network obtained by discarding the rightmost terms along with *student* and *higher education* leads to the star-shaped example in Figure 3.) The network globally splits into three denser regions, and satellite terms lead to a low *cohesion intensity* 0.09 and *cohesion homogeneity* 0.44.

Let us now focus on denser areas in the quest for more semantically cohesive units. The lower part of the network organises around the term *protest* (manifestation) 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 cohesion intensity 0.23 and homogeneity 0.81. Further focusing on *Chile* leads to a 5-node clique with slightly higher cohesion 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 Figure 5), 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 cohesion intensity 0.64 and cohesion homogeneity 0.93 ; again, smaller term and document subsets typically reach higher semantic cohesion. The upper part of the network in Figure 5 gathers documents related to initiation rites in *Grandes Ecoles.* 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 cohesion intensity 0.41 and cohesion homogeneity 0.76. Although cohesion 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 *Grandes Ecoles*, which are all associated with *tradition*, ultimately contained in the set of edges associated with *hazing.*

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

3.3 More on cohesion profiles

We conclude this section and return to cohesion profiles considering the previously discussed examples. We have used cohesion intensity $\frac{\lambda}{|F|}$ and cohesion homogeneity $\frac{\langle 1_T, \gamma \rangle}{\|1_T\| \|\gamma\|}$ as two distinct measures to provide complementary information about the term interaction network. The ex-

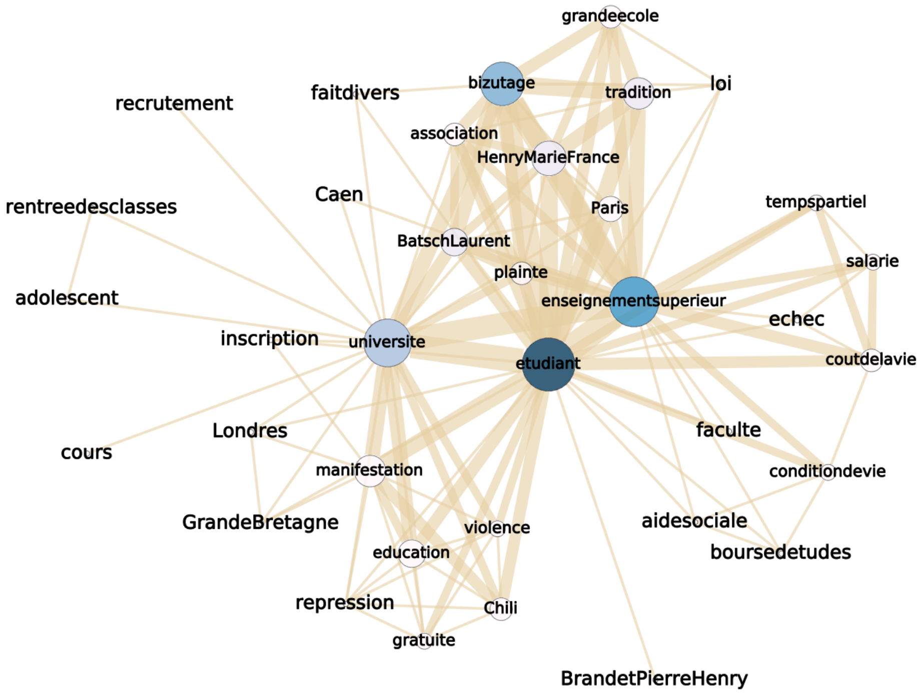


FIGURE 5 – Interaction network induced from a set of documents related to students and universities. This network has a much more intricate topology and twice as many terms as the previous example (Figure 4).

amples 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 cohesion intensity is plotted along the x-axis and cohesion homogeneity is plotted along the y-axis (Figure 6). Any term interaction network would then be plotted as a 2D point $(x, y) = (\frac{\lambda}{|F|}, \frac{\langle 1_T, \gamma \rangle}{\|1_T\| \|\gamma\|})$ in the plane, and we may expect the plot to divide into areas that correspond to network profiles. A simplistic assumption is to expect the 2D plane to subdivide into four more or less rectangular areas, as suggested by the dotted lines. This is far from being acurate, and we may suspect relevant areas to follow more complex patterns. This issue remains to be further studied. We stick to a simplistic, rectangular division of the plane for now.

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 "Radar" 5-node clique in Figure 2 falls into this profile category, as does the "Road safety" upper sub-network considered in the first use case (Figure 4, Section 3.1).

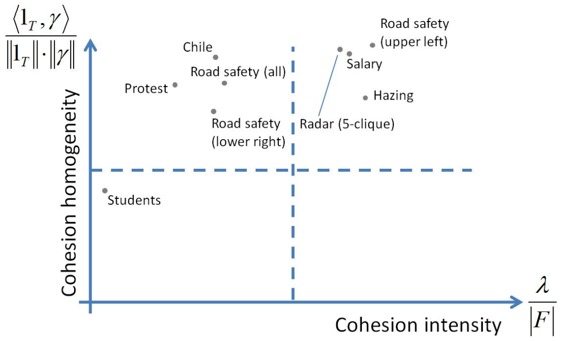


FIGURE 6 – Cohesion profiles can roughly be categorised by combining cohesion intensity and homogeneity and identifying critical areas.

The upper-left 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 categgory are the "Road safety" lower sub-network (Figure 4, Section 3.1), and the *Chile* sub-network with cohesion intensity 0.30 and cohesion homogeneity 0.90 is also part of this category.

The lower-right 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 cohesion intensity 0.44 and cohesion 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 cohesion 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 cohesion intensity of 0.09 and a cohesion homogeneity of 0.44, falls within this category.

Although these areas already provide an interesting grid for evaluating network profiles, more experimentation is needed to assess these prospective categories, determine thresholds defining the profile areas and estimate how they are populated.

4 User study

Section 2 introduced a semantic cohesion index computed using the term interaction network. Semantic cohesion, however, is a notion that is often qualitatively evaluated. We thus felt the need to confront the cohesion index with expert users to assess its relevancy. Four expert documentalists (2 senior, 2 junior) were recruited at the "Multimedia Institute" and asked to conduct an exploratory document collection analysis. Although informal (mainly because we only recruited a small sample of users), the experiment was designed according to a strict protocol. Four different document sets were used ; smaller samples contained between 20 and 40 documents, and larger samples contained between 60 and 80 documents. Each sample contained documents related to events covered in the news (the documents themselves corresponded to TV news excerpts). For each size, the samples showed clear contrasts between cohesion intensity and homogeneity.

Users had access to documents in two different settings. In the first situation, they could view and scroll through a list of documents to freely inspect document titles, content and index terms. In the second situation, users had access to both documents and terms using a graph-based presentation (nodes and links) and could filter documents according to a subset of terms selected from the interaction network. Users had a 5-minute training period to ensure that they could use the interfaces and understood the task they were asked to perform. They were then given random combinations of document samples and interfaces to avoid possible (learning or tiredness) biases. The experiment ended with a questionnaire and face-to-face interview. The experiment took 2 1/2 hours per user on average.

The main goal of the experiment was to get feedback from experts to see whether the cohesion index appropriately defined the terms identified as essential for describing document contents, and if they interacted as the interaction network predicted. Users were asked to rate several aspects of the document groups. We can only report partial results here. Users were asked to do the following :

- evaluate the overall semantic cohesion of each document sample ;
- eliminate "noisy" documents to reinforce the global semantic cohesion within the remaining documents and indicate documents they felt were wrongly indexed (wrong terms) ;
- find, within the sample, documents they felt were more relevant to given queries ;
- tell a story explaining what the document sample contained (which should more or less correspond to the event covered in the news) ;
- express the confidence they had in their analysis (e.g., discarding the *right* documents, recovering the *right* story).

We also compiled answers to the questions users were asked after performing the tasks. According to these answers, users appreciated that the network helped them discriminate between terms. They also liked the conciseness of the network representation and agreed that it gave them a better understanding of the document sample as a whole. Comments from the face-to-face interview revealed that users could develop a good intuition about the document collection from the network shape and use this intuition to identify salient characteristics (e.g., central term sets and outliers). Users agreed that the cohesion indices relate well to the semantic cohesion that they could perceive in a document group. Users also appreciated that lower-cohesion index documents could be easily identified and discarded as 'outliers'. Users' confidence levels clearly improved when using the interaction network to explore the document samples. Finally, users felt that the support gained from the interaction network became more critical as the network and document collection sizes increase.

5 Conclusion

This paper introduced a term interaction network (Section 2) as a device from which term cohesion 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 3) clearly shows the added value brought by the interaction network topology. The network shape is a clear indicator of cohesion profiles, with an obvious archetype profile of an optimally cohesive group as a clique. The examples show how the topology organises terms into areas : some terms are deeply nested into a region, while others act as pivot between regions. Diagram 6 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 [2]. 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 semantically cohesive term and document subsets.

The network profile notion (Sections 2.3 and 3.3) was introduced to distinguish cases where cohesion intensity and homogeneity differ. The informal user study supplemented our use cases and confirmed the network profiles' relevance as a good indicator of salient features in a document collection. Additionally, the interaction network shape is a relevant ingredient for guiding document collection exploration in the quest for semantically cohesive document subsets. Section 3.3 introduced a tentative diagram and categorization of profiles that obviously require more investigation.

The examples used have relatively small sizes. The largest document samples we considered gather hundreds of documents and terms, at most. 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 3, thus leading to less cohesive document and term spaces. Conversely, a close examination of the term interaction network may help identify the core terms from which documents form a cohesive unit.

Our future plans include embedding our prototype into a full-scale application to get proper feedback of its use in realistic situations and ideally performing controlled experiments to objectively establish the benefits of our technique. The network is meant to become a central artefact of a system interface. In that respect, the user interviews helped us identify mandatory interactions for taking full advantage of the network as an exploration device. Being able to automatically identify network profiles would greatly help users. Extending cohesion indices to documents may also be useful when directly exploring or inspecting documents. It would be interesting to compare a document cohesion index to existing ranking indices.

We also plan to examine strategies to automatically identify term and document subsets with optimal (maximal) cohesion intensity and homogeneity. With cohesion indices, we might be able to improve existing document clustering techniques. 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.

Références

[1] A. Aizawa. An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, 39(1) :45–65, 2003.

[2] 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.

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

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

[5] 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.

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

[7] 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.

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

[9] 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.

[10] 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.

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

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

[13] 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.

[14] R. Guimera, S. Mossa, A. Turtschi, and L. A. N. Amaral. The worldwide air transportation network : anomalous centrality, community structure, and cities global roles. *Proceedings*

of the National Academy of Sciences of the United States of America, 102(22) :7794–7799, 2005.

[15] L. Kirchhoff, K. Stanoevska-Slabeva, T. Nicolai, and M. Fleck. Using social network analysis to enhance information retrieval systems. *Procedia - Social and Behavioral Sciences – 4th Conference on Applications of Social Network Analysis (ASNA 2008)*, to appear.

[16] 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.

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

[18] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27 :379–423, 623–656, 1948.

[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] 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.

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

RESEARCH CENTRE
BORDEAUX – SUD-OUEST

351, Cours de la Libération
Bâtiment A 29
33405 Talence Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399