Markov associativities

(The paper has appeared in the Journal of Quantitative Linguistics*, 2005, vol. 12, no. 2, pp. 123-137)*

François Bavaud
Informatique et Méthodes Mathématiques
Université de Lausanne
CH-1015 Lausanne, Switzerland
`francois.bavaud@unil.ch`

Aris Xanthos
Section de Linguistique
Université de Lausanne
CH-1015 Lausanne, Switzerland
`aris.xanthos@unil.ch`

Abstract

Quantifying the concept of co-occurrence and iterated co-occurrence yields indices of similarity between words or between documents. These similarities are associated with a reversible Markov transition matrix, whose formal properties enable us to define euclidean distances, allowing in turn to perform words-documents correspondence analysis as well as words (or documents) classifications at various co-occurrences orders.

1 Introduction

Two objects are associated if they co-occur frequently enough in the same contexts. In the statistical analysis of textual data, objects can be words and contexts can be documents; associativity between words can be defined as proportional to the probability to draw the second word in a document, given that this document contains the first word. One might for instance expect that `théorème` is little associated with `amour` (because few documents co-cite them), `théorème` is strongly associated with `logarithme` (due to the contribution of mathematical documents), and that `amour` and `logarithme` are (almost) not associated.

Associativities defined in this way are closely related to the components of a Markov transition matrix W, giving the probability to reach a word starting from another; we refer to them as *Markov associativities*. By construction, Markov associativities constitute similarity indices obeying well-identified mathematical constraints (symmetry, non-negativity, non-negative definiteness, normalization). They are in principle applicable to any kind of corpus, the choice and organization of which are nevertheless bound to strongly influence the conclusions which may be drawn from this formalism.

Markov associativities can be computed from any words-documents contingency table, giving the number of times n_{jk} word j has occurred in document k. By duality, i.e. by transposing the matrix n_{jk}, the same formalism can be used to define Markov transitions between *documents*, that is Markov documents associativities.

Also, Markov transition matrices can be iterated, yielding higher-order transition matrices (possessing the same stationary distribution). Thus higher-order Markov associativities can be defined in a straightforward way, and capture the idea of higher-order association between objects through "co^r-occurrences", for $r = 1, 2, 3, \ldots$

Markov associativities are non-negative definite, which makes the distances between words euclidean. Words can thus be represented by a configuration of coordinates, the low-dimensional projection of which aims at maximizing the expressed inertia. The resulting procedure amounts to a factorial correspondence analysis (FCA), endowed with familiar words-documents duality properties. Alternatively, hierarchical classification can be performed, yielding classes of similar words, the composition of which generally varies with the order of the associativity under consideration.

2 Notations and formalism

Consider a corpus made of p documents, containing n tokens in total:

- n_{jk} denotes the number of words of type $j = 1, \ldots, m$ occurring in the k-th document $(k = 1, \ldots, p)$
- $n_{j\bullet} := \sum_{k=1}^{p} n_{jk}$ is the absolute frequency of word j
- $n_{\bullet k} := \sum_{j=1}^{m} n_{jk}$ is the size of document k
- $n_{\bullet\bullet} = n = \sum_{j,k} n_{jk}$ is the size of the corpus
- $\pi_j := \frac{n_{j\bullet}}{n}$ is the relative frequency of word j
- $\rho_k := \frac{n_{\bullet k}}{n}$ is the relative size of document k
- $q_{jk} := \frac{n_{jk}\, n}{n_{j\bullet}\, n_{\bullet k}}$ is the associated *independence quotient*, namely the ratio of the observed versus expected count under independence; by construction, $\sum_j \pi_j\, q_{jk} = 1$ and $\sum_k \rho_k\, q_{jk} = 1$.

Words j and j' co-occurring in the same documents $k = 1, \ldots, p$ are associated, and this basic relationship can be quantified by means of an $(m \times m)$ Markov transition matrix $W = (w_{jj'})$ constructed as follows (see figure 1):

1. given a word j, choose a document k with probability $p(k|j) = \frac{n_{jk}}{n_{j\bullet}}$
2. then choose a word j' in document k with probability $p(j'|k) = \frac{n_{j'k}}{n_{\bullet k}}$

The resulting transition matrix reads

$$\begin{aligned} w_{jj'} &= \textstyle\sum_{k=1}^{p} p(k|j)\, p(j'|k) = \sum_{k=1}^{p} \frac{n_{jk}}{n_{j\bullet}} \frac{n_{j'k}}{n_{\bullet k}} \\ &= \textstyle\sum_{k=1}^{p} \rho_k\, q_{jk}\, q_{j'k}\, \pi_{j'} \end{aligned} \tag{1}$$

and enjoys the following properties:

1. $w_{jj'} \geq 0$ and $w_{j\bullet} = 1$, that is W is a Markov transition matrix.

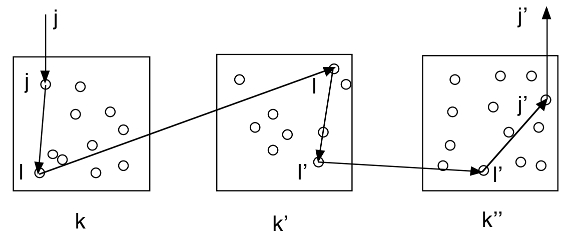

Figure 1: The associativity $s_{jj'}^{(r)}$ of order r (here $r = 3$) is the ratio of the *probability to get the word j' starting from word j* to the *relative frequency of word j'* : first, draw a document k containing word j, pick another word l in k, find another document k' containing l, pick another word l' in k', find another document k'' containing l', and finally pick (or not) word j' in k''.

2. $\sum_{j=1}^{m} \pi_j\, w_{jj'} = \pi_{j'}$, which shows π to be the stationary distribution for W [1].
3. $\pi_j\, w_{jj'} = \pi_{j'}\, w_{j'j}$, that is the Markov chain is *reversible*[2].
4. for $r = 2, 3, \ldots$, the r-th iterate $W^r = (w_{jj'}^{(r)})$ is another transition matrix, defining the *iterated chain of order r*. W^r is also reversible with stationary distribution π, with asymptotic behavior $\lim_{r\to\infty} w_{jj'}^{(r)} = \pi_{j'}$, independently of the initial word or query j.

3 Markov associativities

Definition: the *Markov associativity $s_{jj'}^{(r)}$ of order r between words j and j'* is (fig. 1):

$$s_{jj'}^{(r)} := \frac{w_{jj'}^{(r)}}{\pi_{j'}} \tag{2}$$

Definition (2) makes words associated at order $r = 1$ ($s_{jj'}$ large) if they occur often in the same documents (association of order 1).

[1] this distribution is unique iff n_{jk} is irreducible, namely not degenerate into two or more components - for instance one component containing French words only in French documents and another containing German words only in German documents, with no lexical intersection.

[2] reversibility characterizes here the word-word or document-document association, and does not refer of course to the sequential ordering of words inside documents.

Higher-order associativities $s_{jj'}^{(r)}$ result from an iteration of the process: at order r, words j and j' are considered as more associated than average ($s_{jj'}^{(r)} > 1$) if the probability to obtain a member of the pair from the other is greater than the average probability ($w_{jj'}^{(r)} > \pi_{j'}$, or equivalently $w_{j'j}^{(r)} > \pi_j$). Formally:

1. the $(m \times m)$ associativity matrix $S^{(r)} = (s_{jj'}^{(r)})$ is *non-negative*, *symmetric* (due to the reversibility of $w_{jj'}$) and *normalized* to $\sum_j \pi_j \, s_{jj'}^{(r)} = 1$

2. $S^{(r)} = S\Pi S\Pi \cdots \Pi S$, where $S = S^{(1)}$ and Π is the diagonal matrix containing the π_j. The matrix S, and also $S^{(r)}$, can be shown to be *positive semi-definite* (p.s.d.), i.e. all the associated eigenvalues are non-negative. In particular, $s_{jj'} \leq \sqrt{s_{jj} \, s_{j'j'}}$. Note that $s_{jj'} > s_{jj}$ can occur when j' is a rare word often co-occurring with a frequent word j.

3. Particular cases:
   a) $s_{jj'}^{(0)} = \frac{\delta_{jj'}}{\pi_{j'}}$ b) $s_{jj'}^{(1)} = \sum_k \rho_k \, q_{jk} \, q_{j'k}$
   c) $s_{jj'}^{(\infty)} \equiv 1$.

Associativities $s_{jj}^{(r)}$ thus define *similarity* indices; however, in contrast to a well-established, although little justified tradition, the self-associativity $s_{jj}^{(r)}$ is *not* equal to $s_{\max} = 1$; one finds instead $s_{jj}^{(r)} \geq 1$ with $s_{jj}^{(r)} \neq s_{j'j'}^{(r)}$ in general (this can be justified from the particular form of the transition matrix (1)). By contrast, the weighted average associativity between any word j and all the other words j', *itself included*, is 1: thus in the picture presented here, the more self-associated is a word, the less associated it is with other, *distinct* words[3].

4 Illustrations

Illustrations 1 to 4 are kinds of *Gedankenexperiments*, while illustration 5 constitutes a real example of modest size.

Illustration 1: consider a pair of words (jj') occurring exclusively together, such as (jj') = `cahin - caha`. Then $n_{jk} = n_{j'k}$ for all k, and in particular $q_{jk} = q_{j'k}$ for all k; more precisely, the latter identity holds iff the lexical profiles are proportional, namely $n_{jk} = a \, n_{j'k}$ for all k. Then $s_{jj'} = s_{j'j'} = s_{jj}$, that is j and j' are maximally associated in view of the property $s_{jj'} \leq \sqrt{s_{jj} \, s_{j'j'}}$. Higher-order associativities inherit this property: $s_{jj'}^{(r)} = s_{j'j'}^{(r)} = s_{jj}^{(r)}$.

Illustration 2: two regional synonyms of the standard French `désordre` are j = `brol` (Belgium) and j' = `cheni` (Switzerland). Although chances that j and j' co-occur in the same document are low (for a "normal" corpus), j and j' are likely to be strongly associated with the same words, which results in $s_{jj'}^{(1)} \cong 0$ and $s_{jj'}^{(2)} >> 1$.

Illustration 3: words $\{j\}$ such as `liberté`, `libertés`, `libérer`, `libre`, etc... can be grouped into the same supra-category J, of relative frequency $\pi_J = \sum_{j \in J} \pi_j$ and associated quotient $q_{Jk} = \sum_{j \in J} \frac{\pi_j}{\pi_J} \, q_{jk}$. Also, other words $\{j'\}$ may be grouped into supra-categories J'. The resulting $J = 1, \ldots, M < m$ supra-categories and the associated $(M \times M)$ associativity matrix transform as $s_{JJ'} = \sum_{j \in J} \sum_{j' \in J'} \frac{\pi_j}{\pi_J} \frac{\pi_{j'}}{\pi_{J'}} \, s_{jj'}$, and inherits the properties of non-negativity, symmetry, normalization and p.s.d. However, $s_{JJ'}^{(r)} \neq \sum_{j \in J} \sum_{j' \in J'} \frac{\pi_j}{\pi_J} \frac{\pi_{j'}}{\pi_{J'}} \, s_{jj'}^{(r)}$ in general for $r \geq 2$: words aggregation and Markov iteration do not commute.

Illustration 4: documents $\{k\}$ can also be concatenated into supra-documents $K = 1, \ldots, P < p$, of frequencies $\rho_K = \sum_{k \in K} \rho_j$ and quotients $q_{jK} = \sum_{k \in K} \frac{\rho_k}{\rho_K} \, q_{jk}$. The resulting associativity $\hat{s}_{jj'} = \sum_K \rho_K \, q_{jK} \, q_{j'K}$ is still non-negative, symmetric, normalized and p.s.d. By Jensen's inequality, the diagonal associativities *decrease* under aggregation :

$$s_{jj} = \sum_k \rho_k \, q_{jk}^2 = \sum_K \rho_K \sum_{k \in K} \frac{\rho_k}{\rho_K} \, q_{jk}^2$$

$$\geq \sum_K \rho_K (\sum_{k \in K} \frac{\rho_k}{\rho_K} \, q_{jk})^2 = \sum_K \rho_K \, q_{jK}^2 = \hat{s}_{jj}$$

which shows that, *on average*, off-diagonal associativities increase under aggregation. One gets $\hat{s}_{jj'} \equiv 1$ in the limit of one single document ($p = 1$), and $s_{jj'} = \delta_{jj'}/\pi_j$ in the limit of minimal "one-token documents" ($p = n$).

[3] cf. the behavior of category `DETDEMFS` in illustration 5 below.

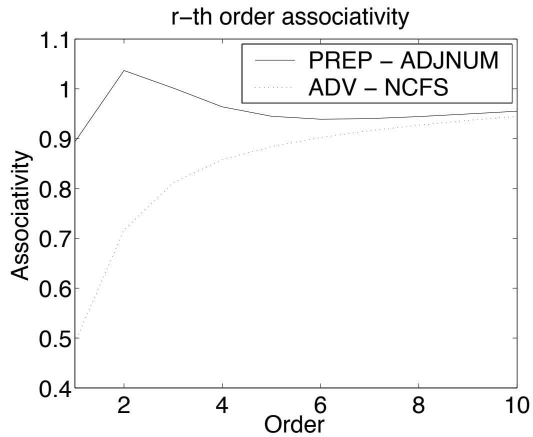

Figure 2: r-th order associativity between unrelated categories.

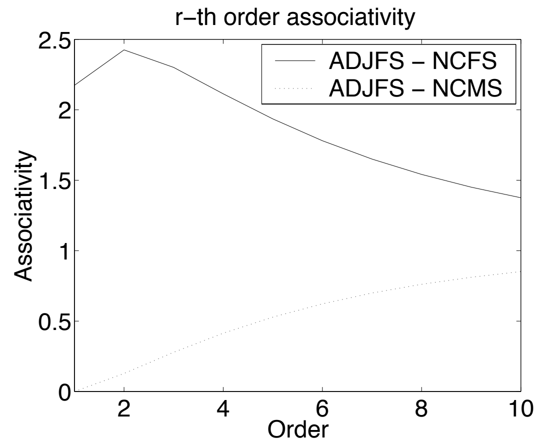

Figure 3: r-th order associativity between syntagmatically related categories.

Illustration 5: in the framework of structural linguistics, it is common to discriminate between *syntagmatic* and *paradigmatic* relationships between linguistic units. The first term refers to units co-occurring within a relevant context, while the second corresponds to units which can be substituted to each other in a given context but cannot occur together[4]. The following illustration shows that, in the domain of syntax, these different relationships yield specific patterns of r-th order associativity. Using the software *CORDIAL Analyseur* developped by the society *Synapse Développement*, we systematically extracted nominal phrases out of a French journalistic corpus[5], replacing the actual words by their syntactic category. After sampling, we obtained a corpus of size $n = 2'914$, containing $p = 1'239$ phrases (documents) and $m = 26$ categories (word types)[6].

After computing the corresponding transition and associativity matrices W^r and S^r at various orders, it turned out that pairs of categories seemed to follow mainly three specific patterns:

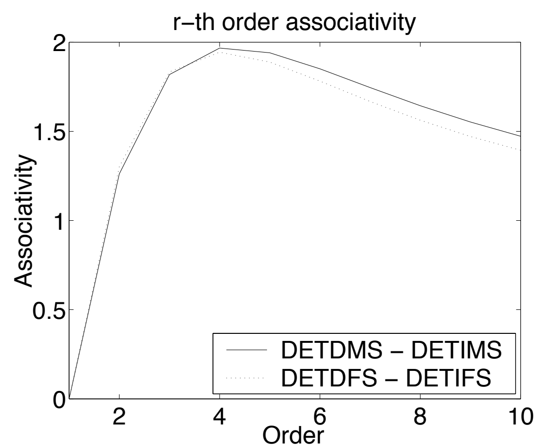

Figure 4: r-th order associativity between paradigmatically related categories.

a) some pairs appear to be only lightly associated at order 1, and tend to exhibit the average associativity of 1 as r grows (possibly crossing that limit, but never in a significant way). For instance, this is the behavior of pairs (`PREP`, `ADJNUM`) or (`ADV`, `NCFS`) (see fig. 2). Elements in such pairs have no particular syntactic relationship together.

b) some other pairs of categories show a high or low first-order associativity, tending to the average associativity as r grows. This behavior characterizes elements with a strong tendency to co-occur (or not) in phrases, like the pairs (`ADJFS`, `NCFS`) or (`ADJFS`, `NCMS`), the second of which violates the rule that noun-adjective groups should possess an uni-

[4] A significant exception to this is the case of coordination.

[5] *La Liberté*, edited in Fribourg, Switzerland.

[6] Key to the abbreviations: `PREP` = preposition, `ADV` = adverb, `NC(M|F)(S|P)` = masculine/feminine singular/plural common noun, `ADJ(M|F)(S|P)` = masculine/feminine singular/plural adjective, `ADJ(S|P)IG` = idem, gender-invariant, `DET(I|D|DEM|POSS)(M|F)S` = indefinite/definite/demonstrative/possessive masculine/feminine singular article, `DET(I|D|DEM|POSS)(S|P)IG` = idem, singular/plural gender-invariant.

fied gender in French (see fig. 3).

c) the last case is that of mutually exclusive elements but liable to "co-co-occur" within the same contexts. Their associativity is minimal for $r = 1$, as they never occur in the same phrase, but it goes significantly beyond the average for $r \geq 2$ before regressing to it for higher orders. Pairs (DETDMS, DETIMS) and (DETDFS, DETIFS) are prototypical examples of this (see fig. 4).

5 Markov dissimilarities: FCA and classification

Associativities $s^{(r)}_{jj'}$ are positive semi-definite, and play the role of the "scalar product matrix" in the *classical multidimensional scaling* problem (see e.g. Schoenberg (1935) or Gower (1982)). Following the latter, we construct euclidean representable dissimilarities $D^{(r)}_{jj'}$ of order r as

$$D^{(r)}_{jj'} := s^{(r)}_{jj} + s^{(r)}_{j'j'} - 2 s^{(r)}_{jj'} = \frac{w^{(r)}_{jj}}{\pi_j} + \frac{w^{(r)}_{j'j'}}{\pi_{j'}} - 2\frac{w^{(r)}_{jj'}}{\pi_{j'}} \quad (3)$$

The weighted average dissimilarity between all pairs of words is the *inertia of order r* defined as

$$I^{(r)} := \frac{1}{2}\sum_{jj'} \pi_j \, \pi_{j'} \, D^{(r)}_{jj'}$$
$$= \sum_j \pi_j \, s^{(r)}_{jj} - \sum_{jj'} \pi_j \, \pi_{j'} \, s^{(r)}_{jj'} = \sum_j w^{(r)}_{jj} - 1$$

Hence, the higher the probability of getting the same word (that is the higher the average self-associativity), the higher the corresponding inertia. Particular cases are $I^{(0)} = m - 1$, $I^{(1)} = \sum_j \pi_j \, s_{jj} - 1$, $I^{(2)} = \sum_{jj'} \pi_j \, \pi_{j'} \, s_{jj'} - 1$ and $I^{(\infty)} = 0$. The inertia of order $r = 1$ is nothing but the chi-square (per count) associated to the words-documents contingency table (n_{jk}) (see also Bavaud (2002)):

$$I^{(1)} = \sum_j \pi_j \, s_{jj} - 1 = \sum_{jk} \pi_j \, \rho_k \, q^2_{jk} - 1$$
$$= \frac{1}{n}\sum_{jk} \frac{(n_{jk} - n \, \pi_j \, \rho_k)^2}{n \, \pi_j \, \rho_k} = \frac{\chi^2}{n}$$

Factorial correspondence analysis (FCA) aims at representing words $j = 1, \ldots, m$ as points $x_{j\alpha}$

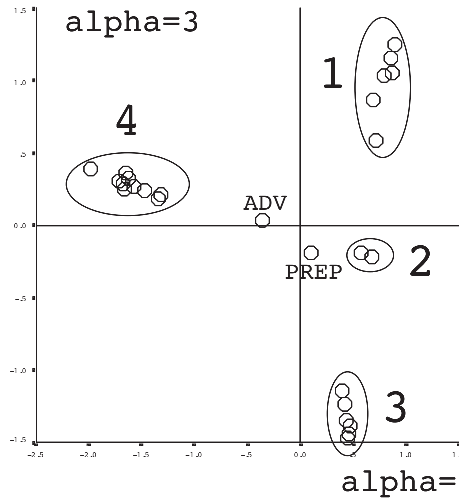

Figure 5: FCA scores for words. *Singular-plural* factor $\alpha = 2$ opposes cluster 1 (ADJMS, DETDEMMS, DETDMS, DETIMS, DETPOSSMS, NCMS), cluster 2 (ADJSIG, DETPOSSSIG) and cluster 3 (ADJFS, DETDEMFS, DETDFS, DETIFS, DETPOSSFS, NCFS) to cluster 4 (ADJFP, ADJMP, ADJNUM, ADJPIG, DETDEMPIG, DETDPIG, DETIPIG, DETPOSSPIG, NCFP, NCMP). *Masculine-feminine* factor $\alpha = 3$ opposes cluster 1 to cluster 3.

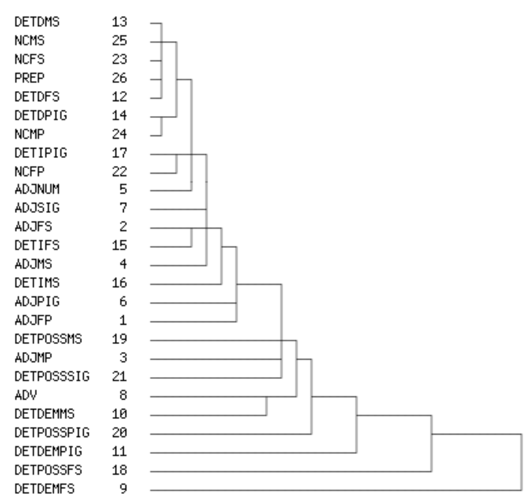

Figure 6: classification on $D^{(1)}_{jj'}$ as defined in (3); as shown in figure 8, Ward classification on $D^{(3)}_{jj'}$ matches more closely the FCA than does the present classification.

such that a maximum part of inertia $I^{(1)}$ is expressed by the first dimensions $\alpha = 1, 2, \ldots$ (see e.g. Greenacre (1984) or Lebart et al. (1995)). The resulting coordinates $\{x_{j\alpha}\}$ constitutes a low-dimensional, factorial representation of words, in contrast to the high-dimensional, direct representation $\{x_{jl}\}$ introduced above.

Higher-order FCA, generalizing the ordinary FCA of order one, can be constructed as follows: consider the spectral decomposition $C^{(r)} = U^{(r)}\,\Lambda^{(r)}\,(U^{(r)})'$ (with $U^{(r)}$ orthogonal and $\Lambda^{(r)}$ diagonal with decreasingly ordered values) of the symmetric $(m \times m)$ matrix $C^{(r)} = (c^{(r)}_{jj'})$ defined as $c^{(r)}_{jj'} := \sqrt{\pi_j}\; w^{(r)}_{jj'}/\sqrt{\pi_{j'}}$; identity $C^{(r)} = C^r$ entails $U^{(r)} = U = (u_{j\alpha})$ (independently of r) and $\Lambda^{(r)} = \Lambda^r$ with diagonal components λ^r_α. The searched for words coordinates are $x^{(r)}_{j\alpha} := \frac{\sqrt{\lambda^r_\alpha}}{\sqrt{\pi_j}}\, u_{j\alpha}$, obeying $\sum_\alpha (x^{(r)}_{j\alpha} - x^{(r)}_{j'\alpha})^2 = D^{(r)}_{jj'}$ as requested[7].

Also, $I^{(r)} = \sum_{\alpha \geq 2} \lambda^r_\alpha$, which shows the first non-trivial dimensions $\alpha = 2, 3 \ldots$ to express a maximum part of the projected inertia $I^{(r)}$ ($\lambda_1 = 1$ corresponds to the trivial eigenvalue) (see fig. 5). As in ordinary FCA, duality enables the same eigen-structure to generate the higher-order factorial representation of *documents* coordinates $y^{(r)}_{k\alpha}$.

Finally, a *classification* of words can be performed: figures 6, 7 and 8 show the results of hierarchical Ward classifications applied on (3) with $r = 1, 2, 3$ respectively. Cutting the dendrograms at some height h (here represented horizontally in arbitrary units) aggregates the m words j into $M \leq m$ groups J, with group coordinates $x^{(r)}_{Jl} := \sum_{j \in J} \frac{\pi_j}{\pi_J}\, x^{(r)}_{jl}$; inertia of order r decomposes into a between- and a within-group contribution:

$$I^{(r)} = \frac{1}{2}\sum_{jj'} \pi_j\,\pi_{j'}\,D^{(r)}_{jj'} = \frac{1}{2}\sum_{JJ'} \pi_J\,\pi_{J'}\,D^{(r)}_{JJ'} + \\ + \sum_J \pi_J \sum_{j \in J} \frac{\pi_j}{\pi_J}\, D^{(r)}_{jJ} =: I^{(r)}_B + I^{(r)}_W$$

[7]Proof: $\sum_\alpha (x^{(r)}_{j\alpha} - x^{(r)}_{j'\alpha})^2 = \sum_\alpha \lambda^r_\alpha\, (\frac{u_{j\alpha}}{\sqrt{\pi_j}} - \frac{u_{j'\alpha}}{\sqrt{\pi_{j'}}})^2$
$= \frac{c^{(r)}_{jj}}{\pi_j} + \frac{c^{(r)}_{j'j'}}{\pi_{j'}} - 2\frac{c^{(r)}_{jj'}}{\sqrt{\pi_j}\sqrt{\pi_{j'}}} = \frac{w^{(r)}_{jj}}{\pi_j} + \frac{w^{(r)}_{j'j'}}{\pi_{j'}} - 2\frac{w^{(r)}_{jj'}}{\pi_{j'}}$
$= s^{(r)}_{jj} + s^{(r)}_{j'j'} - 2s^{(r)}_{jj'} = D^{(r)}_{jj'}$.

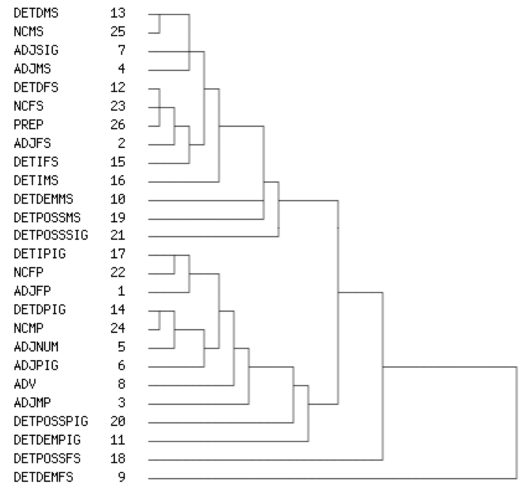

Figure 7: Ward classification on $D^{(2)}_{jj'}$.

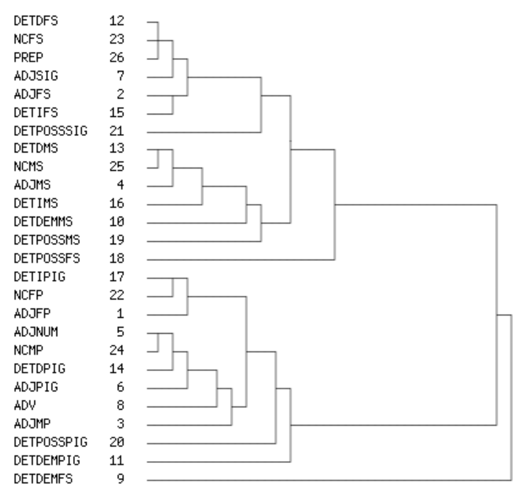

Figure 8: Ward classification on $D^{(3)}_{jj'}$.

where $D^{(r)}_{jJ} := \sum_l (x^{(r)}_{jl} - x^{(r)}_{Jl})^2$ is the *word-group* dissimilarity $D^{(r)}_{jJ} := \sum_l (x^{(r)}_{jl} - x^{(r)}_{Jl})^2$ and $D^{(r)}_{JJ'} := \sum_l (x^{(r)}_{Jl} - x^{(r)}_{J'l})^2$ the *group-group* dissimilarity. Under aggregation $J, J' \to [J \cup J']$, the intra-group inertia $I^{(r)}_W$ increases of $\Delta I^{(r)}_W = \frac{\pi_J \, \pi_{J'}}{\pi_J + \pi_{J'}} D^{(r)}_{JJ'}$. Aggregating groups in a way which minimizes this increase (as in figures 6, 7 and 8) amounts to Ward clustering algorithm (see e.g. Lebart et al. (1995)).

Interestingly enough, changing the order $r \to r'$ transforms the FCA representation into another FCA representation which is pretty close to the former since the eigenvalues solely are altered as $\lambda^r_\alpha \to \lambda^{r'}_\alpha$; by contrast, the associated classification can be altered fairly more substantially, as attested by figures 6, 7 and 8.

6 Conclusion and further developments

The present work has explored a few *formal* properties of the concept of *associativity of order* r, demonstrating how it can be statistically founded and used in a classical data-analytical framework.

In the vector space representation of information retrieval (IR) (see e.g. Slaton and Buckley (1988); Besançon et al. (1999)), document-document similarities are typically defined as $\tilde{\sigma}_{kk'} = \frac{(a_k, a_{k'})}{\sqrt{(a_k, a_k)(a_{k'}, a_{k'})}}$ where $(a_k, a_{k'}) := \sum_j a_{kj} \, a_{k'j}$ and a_{kj} is the vector of terms weigths associated with document k, with

$$a_{kj} = \begin{cases} (1 + \log n_{jk}) \log \frac{p}{\sharp\{k|n_{jk}>0\}} & \text{if } n_{jk} > 0 \\ 0 & \text{otherwise} \end{cases}$$

By contrast, first-order Markov document-document similarities (1) express as $\tilde{s}_{kk'} = (b_k, b_{k'})$ where $b_{kj} = \sqrt{\frac{n}{n_{j\bullet}} \frac{n_{jk}}{n_{\bullet k}}}$. Contrarily to $\tilde{\sigma}_{kk'}$, the associativity $\tilde{s}_{kk'}$ is invariant under the aggregation of words possessing identical profiles, as does the generalized family $b_{kj} = \sqrt{\frac{n_{j\bullet}}{n}} \; f(\frac{n_{jk} \, n}{n_{j\bullet} n_{\bullet k}})$ (Bavaud 2002). In that respect, Markov associativities could play the role of reference similarities, endowed with appealing formal properties, to which the various *tf-idf* weighting schemes proposed and evaluated in the literature might be compared.

Also, the present formalism could be further developped by considering *fuzzy* memberships and associativities (Bavaud 2004), or by incorporating work on *probabilistic latent semantic analysis* (Hofmann 1999), postulating conditional probability of the form $p(j|k) = \sum_z p(j|z) \; p(z|k)$ where z indexes latent classes. Others extensions implying non-linear distortions of the distances conserving the euclidean property, trade-off between orders by using chain mixtures, and special documents definitions enabling links with the n-grams formalism are currently under investigation.

Although we are confident about the formal strength of our formalism, which we judge as sound and statistically founded (and obviously not restricted to textual data), results on large-scale and systematic empirical performance of IR systems based upon the present formalism are presently yet missing. This state of things should be remedied in priority: at the time being, the question of whether our formalism will perform better in practice that another system based upon somewhat ad hoc assumptions remains open.

Acknowledgements

Thanks to M. Rajman and J.-C. Chappelier for stimulating discussions, and to N. Jufer and S. Durrer for their textual data.

References

François Bavaud. 2002. *Quotient Dissimilarities, Euclidean Embeddability, and Huygens' Weak Principle.* In K. Jajuga and al. (Eds.): *Classification, Clustering, and Data Analysis.* 195-202. Springer.

François Bavaud. 2004. *On the Comparison and Representation of Fuzzy Partitions.* To appear in *Student*.

John C. Gower. 1982. *Euclidean distance geometry*. The Mathematical Scientist, 7:1-14

Michael J. Greenacre. 1984. *Theory and applications of Correspondence analysis*. Academic Press, New York.

Thomas Hofmann. 1999. *Probabilistic Latent Semantic Indexing.* In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99).* Berkley, USA. 50-57.

Ludovic Lebart, Alain Morineau, and Marie Piron. 1995. *Statistique exploratoire multidimensionelle*. Dunod, Paris.

Romaric Besançon, Martin Rajman and Jean-Cédric Chappelier. 1999. *Textual Similarities based on a Distributional Approach*. In *Proceedings of the Tenth International Workshop on Database and Expert Systems Applications (DEXA99)*. Firenze, Italy. 180-184.

Gerard Slaton and Chris Buckley. 1988. *Term weighting approaches in automatic text retrieval*. Information Processing and Management, 24:513-523

Isaac J. Schoenberg. 1935. *Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espaces vectoriels distancés applicables vectoriellement sur l'espace de Hilbert"*. Annals of Mathematics, 36:724-732