A Second-order Conglomerate Construction and Correlation Studies in a Conglomerate Framework

Ronald Rousseau

ronald.rousseau@khbo.be
KHBO (Association K.U.Leuven), Department of Industrial Sciences and Technology, Zeedijk 101, B-8400 Ostend (Belgium)

Introduction

A conglomerate is a framework for informetric (and other) research (Rousseau, 2005). It consists of two collections, a source collection and a pool, and two mappings (a precise definition is provided in the next section). Main elements in the conglomerate construction are the conglomerate ratio, a kind of average, defined by analogy to the journal impact factor, and three lists drawn for analyses of Zipf or Lotka type. It is shown in this poster how a second order conglomerate can be constructed from a given conglomerate and a partition of its source collection. It is further shown how correlation analysis fits into the conglomerate framework. Constructions are illustrated using impact factors and university research group production.

Conglomerates: Definition

A conglomerate consists of two collections and two mappings. The first collection is a finite set, denoted as S, and called the source collection. Its elements are called sources. The second collection is called the pool. It is not necessarily finite, but in practical applications it will always be finite. Further a mapping f is given from S to 2^P, the set of all subsets of P. For each s $\in$ S, f(s) is a subset of P, called the item-set of s. The union of all p in P belonging to at least one item-set is called the item collection, denoted as I $\subset$ P. The map f itself is called the source-item map.

Next, each set f(s) is mapped to a number, called the magnitude of this set. This mapping is denoted as m and maps f(s) $\in 2^P$ to m(f(s)) $\in \mathbb{R}^+$. The mapping itself is called the magnitude function. Often, but not always, m will be the counting measure which maps f(s) to the number of elements in f(s). So we conclude that *a conglomerate is a 4-tuple* $\boldsymbol{C}$ *= (S,P,f,m)* consisting of a source collection S, a pool P, a mapping $f : S \to 2^P$, and a mapping $m : D \subset 2^P \to \mathbb{R}^+$ where $\mathrm{I} = \bigcup_{s \in S} f(s) \subset D$.

These steps lead to a first important topic in informetric research, namely the ratio of the sum of all magnitudes of item-sets, and the number of elements in the source collection, termed the conglomerate ratio, and denoted as μ_C.

$$\text{Conglomerate ratio} = \mu_c = \frac{\sum_{s \in S} m(f(s))}{\#(\text{source collection})}$$

Finally, the source-item relation of a conglomerate leads to three lists. The first one just consists of all sources and the magnitude of their corresponding item sets. This is called the conglomerate list. The second list is the same as the first one, but sources are ranked according to the magnitude of their corresponding item-sets. We will refer to this list as a Zipf list. The first list can also, if desired, be rewritten in size-frequency form, leading to a third list associated with the source-item relation of a conglomerate.

Examples of objects of study that can be defined using a conglomerate framework include: impact factors, author production data, collaboration data, web impact factors, bestsellers lists, diffusion factors and election results, see (Rousseau, 2005).

Construction of a Second Order Conglomerate

In this section we present a construction that will make it possible to construct a new, higher order conglomerate, from a given one. Its best-known application is the construction of the conglomerate related to the average impact factor of a group of journals starting from the conglomerate leading to the global impact factor.

Construction

Let **C** = (S,P,f,m) be a conglomerate as defined above. In this context it is referred to as the basis conglomerate. Let $(S_j)_{j=1,\ldots,n}$ be a partition of S. This means that none of the sets S_j is empty, each intersection of two is empty, and their union is S. In symbols:

$$\forall j : S_j \neq \varnothing;\ \forall i, j, i \neq j : S_i \cap S_j = \varnothing;\ \bigcup_{j=1}^{n} S_j = S .$$

We construct a new conglomerate **C'** = (T, Q, F, M) as follows. T is the set $\{S_1, S_2, \ldots, S_n\}$, Q is the product set $P \times S$, F: T $\to$ 2^Q: $S_j \to$

$\bigcup_{s\in S_j}\{(f(s),s)\} \subset 2^P \times S_j \subset 2^P \times S \subset 2^{P\times S}$. Finally $M : D \subset 2^{P\times S} \to \mathbb{R}^+$ is defined on D, the union of all F(S_j), as: $M(F(S_j)) = \frac{\sum_{s\in S_j} m(f(s))}{\# S_j}$. The value of M is nothing but the conglomerate ratio of the original conglomerate restricted to the set S_j.

Two Examples

Example 1. Let S be the set of all journal articles published in the journals $S_1,\ldots S_n$, during the years Y-1 and Y-2; P is the set of all references of all articles included in the Web of Science in the year Y. The map f maps each article to the set of all its citations. Finally, m is just the counting measure. Then the conglomerate ratio of **C** = (S,P,f,m) is the global impact factor of this set of journals. This is the basis conglomerate.

S is clearly partitioned into the sets of articles published in journals $S_1,\ldots,S_n$. The source collection for the second order conglomerate is the set of all journals. The function F maps each journal to the disjoint union of each of its articles' citations. Finally F(M(S_j)) is the Garfield-Sher impact factor of journal S_j in the year Y. The conglomerate ratio of the second order conglomerate is just the average impact factor of the set of journals under study. The list resulting from this construction is the list of all Garfield-Sher impact factors of the journals.

Example 2. As a second example we consider the group of all scientists of a university department. Each scientist, s, is mapped to the set, f(s), of articles he/she (co)authored over a given period of time. When m is the counting measure then m(f(s)) is the number of articles published by scientist s. The conglomerate ratio is the average number of articles published by scientists of this department.

Next, we partition scientists according to the research group within the department, to which they belong, and apply the construction of the second-order conglomerate. Denoting these groups as $g_1,\ldots,g_n$, we see that F(g_j) is nothing but the disjoint union of sets of articles (co)authored by each of the members of research group g_j. Finally, M(F(g_j)) is the average number of articles (co)authored by members of this research group. The conglomerate ratio of this second order conglomerate is the average (taken over all research groups) of the average production of each research group. The conglomerate list consists of the research groups of this department and the average number of published articles.

Correlation Analysis

When two conglomerates have the same source collection this leads to two corresponding sets of lists. Studying relations between these lists can be done using correlation analysis. Pearson correlations are used for the first type of lists, while Spearman correlation is used for the ranked forms. In this way correlation analysis finds a natural place in the conglomerate framework. Some examples will clarify what we mean to say.

Example 3. Consider first a conglomerate having a group of articles as source collection. Each article, s, is mapped by a map f_1 to the set of citations it received over a fixed period of time. All citations come from the pool of ISI-covered journals. The magnitude map, m_1, is just the number of received citations. Considering the same source collection, one can map each article, s, by a mapping denoted as f_2, to the first article (in the pool) citing it. The magnitude function m_2 maps f_2(s) to the time between the publication of s and its first citation. It is now interesting to study the correlation between the time to first citation and the total number of citations. Such a study naturally finds its place in a conglomerate framework. The next example uses the second order conglomerate construction.

Example 4. Studying the correlation between the Garfield-Sher impact factor and the immediacy index. A construction of a second order conglomerate applied to a set of n journals may lead to a list of standard impact factors, as shown above. In a similar way, another second order conglomerate construction can be applied to the same n journals with as original pool (P) the set of all references of all articles included in the Web of Science in the years Y-1 and Y-2. Mapping each article s to the set of citations in the year of publication gives f(s). The mapping m is again the counting measure. Then the conglomerate ratio of the basis conglomerate is the global immediacy index of this set of journals (covering two years!). Strictly speaking this is not an immediacy index but a kind of generalized impact factor, as introduced by Frandsen and Rousseau (2005). We next partition S into 2n subsets according to journal title and publication year. Then the second order conglomerate construction defines M(F(S_j(Y-k))) as the immediacy index of journal S_j in the year Y-k (k = 1,2). We may now perform a correlation analysis between the list of impact factors, the list of immediacy indices for the year Y-1, and the list of immediacy indices for the year Y-2.

References

Frandsen, T. F. & Rousseau, R. (2005). Article impact calculated over arbitrary periods. *Journal of the American Society for Information Science and Technology*, 56, 58-62.

Rousseau, R. (2005). Conglomerates as a general framework for informetric research. *Information Processing and Management (to appear)*.