A Similarity Measure to Generalize Attributes

Rostand S. Kuitché[1], Romuald E. A. Temgoua[2], and Léonard Kwuida[3]

[1] Université de Yaoundé I, Département des Mathematiques,
BP 812 Yaoundé, Cameroun
kuitcher@yahoo.com
[2] Université de Yaoundé I, École Normale Supérieure,
BP 47 Yaoundé, Cameroun
retemgoua@gmail.com
[3] Bern University of Applied Sciences,
Brückenstrasse 73, 3005 Bern, Suisse
leonard.kwuida@bfh.ch

Abstract. Formal Concept Analysis (FCA) plays a crucial role in various domains, especially in qualitative data analysis. Here knowledge are extracted from an information system in form of clusters (forming a concept lattice) or in form of rules (implications basis). The number of extracted pieces of information can grow very fast. To control the number of cluster, one possibility is to put some attributes together to get a new attribute called a generalized attribute. However, generalizing does not always lead to the expected results: the number of concepts can even exponentially increase after generalizing two attributes [7,8]. A natural question is whether there is a similarity measure, (possibly cheap and fast to compute), that is compatible with generalizing attributes: i.e. if m_1, m_2 are **more similar** than m_3, m_4, then putting m_1, m_2 together should not lead to more concepts as putting m_3, m_4 together. This paper is an attempt to answer this question.

Keywords: Formal Concept Analysis; Generalizing Attributes; Similarity Measures.

1 Introduction

In Formal Concept Analysis (FCA), a **formal context** is a binary relation (G, M, I) that models an elementary information system, whereby G is the set of objects, M the set of attributes and $I \subseteq G \times M$ the incidence relation. To extract knowledge from such an elementary information system, one possibility is to get clusters of objects and/or attributes by grouping together those sharing the same characteristics. These pairs, called **concepts**, were formalized by Rudolf Wille [16]. For $A \subseteq G$ and $B \subseteq M$ we set

$$\begin{aligned} A' &= \{m \in M \mid g \,\mathrm{I}\, m \text{ for all } g \in A\} \text{ and} \\ B' &= \{g \in M \mid g \,\mathrm{I}\, m \text{ for all } m \in B\}. \end{aligned}$$

A **concept** is a pair (A, B) such that $A' = B$ and $B' = A$. A is called **extent** and B **intent** of the concept (A, B). The set of concepts of a context $\mathbb{K} := (G, M, \mathrm{I})$ is ordered by the relation $\quad (A, B) \leq (C, D) :\iff A \subseteq C$, and forms a lattice, denoted by $\mathfrak{B}(\mathbb{K})$ and called **concept lattice** of $\mathbb{K}$. To control the size of concept lattices, many methods have been suggested: decomposition [18,19,17], iceberg lattices [14] α-Galois lattices [15], fault tolerant patterns [3], closure or kernel operators and/or approximation [6]. In [7] the authors consider putting together some attributes to get a generalized attribute. Doing this one has to decide when an object satisfies a (new) generalized attribute. They discuss several scenarios among which the following, called $\exists$-generalization:

> an object $g \in G$ satisfies a generalized attribute $s \subseteq M$ if g satisfies at least one of the attributes in s. i.e. $s' = \bigcup\{m' \mid m \in s\}$.

In the rest of this contribution, we will simply say **generalization** to mean $\exists$-generalization. By generalizing (i.e putting together some attributes) we reduce the number of attributes and hope to also reduce the size of the concept lattice. Unfortunately this is not always the case. In [8] the authors provide some examples where the size increases exponentially after generalizing two attributes and also give the maximal increase.

In [1,5], the authors discuss similarity measures on concepts, and even on lattices. For our purpose, we need a measure of similarity on attributes such that if m_1, m_2 are more similar than m_3, m_4, then generalizing m_1, m_2 should not lead to more concepts as generalizing m_3, m_4. We say that such a similarity measure is **compatible with the generalization**. Given a set M of attributes, a **similarity measure** on M is defined as a function $S : M \times M \to \mathbb{R}$ such that for all m_1, m_2 in M,

(i) $S(m_1, m_2) \geq 0$, **positivity**
(ii) $S(m_1, m_2) = S(m_2, m_1)$ **symmetry**
(iii) $S(m_1, m_1) \geq S(m_1, m_2)$ **maximality**

If in addition $S(m_1, m_2) \leq 1$, we say that S is **normalized**. Similarity measures aim at quantifying to which extent two attributes resemble each other. Getting a similarity measure compatible with the generalization will be a valuable tool in preprocessing and will warn the data analyst on possible lost or gain when generalizing.

The rest of the paper is organized as follows: In Section 2, we investigate the existing similarity measures that we found in the literature. In Section 3, we give a new similarity measure that characterize the pairs of attributes which can increase the size of the concept lattice after generalizing. Section 4 exposes an example on lexicographic data and Section 5 concludes the paper.

2 Test of Existing Similarity Measures in ∃-Generalization

Similarity and dissimilarity measures play a key role in pattern analysis problems such as classification, clustering, etc. Ever since *Pearson* proposed a coefficient of correlation in 1896, numerous similarity measures and distance have been proposed in various fields. These measures can be grouped into tree main types, depending of the data on which they are used:

Correlation coefficients: They are often used in data to compare variables with qualitative characters subdivided in more than two states.

Distance similarity coefficients: They are generally used in data with pure quantitative variables. In most cases, for quantitative data, the similarity between two taxa is expressed as a function of their distance in a dimensional space whose coordinates are the characters.

Coefficients of association: They are often used in data with presence-absence characters or in data with individuals having qualitative characters subdivided into two states.

There are two subsets of coefficients of association: those that only depend on characteristics present in at least one of the taxa compared, but are independent of the attributes absent in both taxa (denoted by type 1), and those that also take into account the attributes absent in both taxa (denoted by type 2). Those measures use

- a as the number of cases where the two variables occur together in a sample,
- d as the number of cases where none of the two attributes occur in a sample,
- b as the number of cases in which only the first variable occur, and
- c as the number of cases where only the second variable occur.

One of the most important similarity measure of type 1 is the **Jaccard measure** $\left(\frac{a}{a+b+c}\right)$, proposed in order to classify ecological species. Also in the ecological field, the **Dice coefficient of association** $\left(\frac{2a}{2a+b+c}\right)$ aims at quantifying the extent to which two different species are associated in a biotope, the **Sorensen coefficient of association** $\left(\frac{4a}{4a+b+c}\right)$ and the **Anderberg coefficient of association** $\left(\frac{8a}{8a+b+c}\right)$ are of the same type. The **Sneath and Sokal 2** similarity coefficient $\left(\frac{\frac{1}{2}a}{\frac{1}{2}a+b+c}\right)$, put in place in order to compare organisms in numerical taxonomy, the **Kulczynski similarity** measure $\left(\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})\right)$ and the **Ochiai similarity** measure $(\frac{a}{\sqrt{(a+b)(a+c)}})$ are also from this first type.

The most used similarity coefficient of the second type is the **Sokal and Michener** coefficient of association $\left(\frac{a+d}{a+d+b+c}\right)$, also called the **simple matching coefficient**, put in place to express the similarity between two species of bees. Moreover, the **Rogers and Tanimoto similarity measure** $(\frac{\frac{1}{2}(a+d)}{\frac{1}{2}(a+d)+b+c})$ whose aim was to compare species of plants in the ecological field, the **Sokal and Sneath 1** similarity coefficient $(\frac{2(a+d)}{2(a+d)+b+c})$ was defined to make comparison in numerical taxonomy and the **Russels and Rao** similarity measure $(\frac{a}{a+d+b+c})$ put in place with the aim of showing resemblance between species of *anopheline* *larvae,* are included in this type. Same are the **Yule and Kendall similarity coefficients** $\left(\frac{ad}{ad+bc}\right)$, often used in the statistical field. Some of the above similarity measures can be found in [5].

Regarding the definitions of the above kinds of similarity measures, only the coefficients of association suitable to formal contexts, since formal contexts are data with presence-absence characters. We will investigate the impact of these coefficients of association on a special pair of attributes in some formal contexts. The objective is to show that these similarity measures are not helpful in finding whether their generalization increases the size of the lattice or not.

Our first example is an arbitrary formal context (G, M, I) containing two attributes $x, y \in M$ such that $x' \subseteq y'$ and $|x' \cap y'| = 1$. Then $|x' \setminus y'| = 0$ and the generalization of the attributes x and y does not increase the size of the lattice. Choosing $|y' \setminus x'| = 20$ and $|G \setminus (x' \cup y')| = 1$ yields $a = |x' \cap y'| = 1$, $b = |x' \setminus y'| = 0$, $c = |y' \setminus x'| = 20$ and $d = |G \setminus (x' \cup y')| = 1$. For the coefficient of association of type 1 with Jaccard (Jc), Dice (Di), Sorensen (So), Anderberg (An), Sneath and Sokal 2 (SS_2), Kulczynski (Ku) and Orchiai (Orch), and the coefficient of association of type 2 with Sokal and Michener (SM), Rogers and Tanimoto (RT), Sneath and Sokal 1 (SS_1) and Russel and Rao (RR), we get the table below for $s(x, y)$:

Jc	Di	So	An	SS_2	Ku	Orch	SM	RT	SS1	RR
0,05	0,09	0,17	0,29	0,02	0,52	0,22	0,09	0,05	0,17	0,05

The table above shows that with almost all these measures, the similarity measured between the attributes x and y is very low, despite the fact that their generalization does not increase the size of the lattice.

Our second example is the formal context $\mathbb{K}_6 := (S_6 \cup \{g_1\}, S_6 \cup \{m_1, m_2\}, \mathrm{I})$ below, with $S_6 = \{1, 2, 3, 4, 5, 6\}$.

$\mathbb{K}_6$	1	2	3	4	5	6	m_1	m_2
1		×	×	×	×	×	×	
2	×		×	×	×	×	×	×
3	×	×		×	×	×	×	×
4	×	×	×		×	×	×	×
5	×	×	×	×		×	×	×
6	×	×	×	×	×			×
g_1	×	×	×	×	×	×		

We observe that $|m_1' \cap m_2'| = 4$, $|m_1' \backslash m_2'| = 1$ and $|m_2' \backslash m_1'| = 1$. Putting together the attributes m_1 and m_2 by a $\exists$-generalization increases the size of the lattice by 16. The following table shows the measures of type 1 and type 2 between the attribute m_1 and any other attribute i. All the similarity measures of the

	Jc	Di	So	An	SS$_2$	Ku	Orch	SM	RT	SS1	RR
$i \in S_5$	0,57	0,80	0,89	0,94	0,50	0,80	0,80	0,71	0,56	0,83	0,57
$i = 6$	0,83	0,91	0,95	0,97	0,71	0,92	0,91	0,75	0,75	0,92	0,71
$i = m_2$	0,67	0,80	0,89	0,94	0,50	0,80	0,80	0,71	0,56	0,83	0,57

two types show that the attribute m_1 is more similar to m_2 than to any other attribute $i \in S_6$ (apart from $i = 6$); But putting m_1 and m_2 together increases the size of the lattice. We can conclude that these similarity measures are not compatible with the ∃-generalization. We are actually looking for a measure on attributes that will flag pairs of attributes as **less similar** when putting these together increases the size of the concept lattice.

3 A Similarity Measure Compatible with ∃-Generalization

In this section we define a similarity measure on attributes which is compatible with the existential generalization. This generalization means that from an attribute reduced context $\mathbb{K} := (G, M, \mathrm{I})$, two attributes a, b are removed and replaced with an attribute s defined by $s' = a' \cup b'$. We set $M_0 := M \setminus \{a, b\}$ and

$$\begin{aligned} \mathbb{K}_{00} :=& (G, M_0, \mathrm{I} \cap (G \times M_0)), && \text{(removing } a, b \text{ from } \mathbb{K}) \\ \mathbb{K}_{0s} :=& (G, M_0 \mathbin{\dot\cup} \{s\}, I_0^s), && \text{(adding } s \text{ to } \mathbb{K}_{00}) \end{aligned}$$

where $I_0^s := (\mathrm{I} \cap (G \times M_0)) \cup \{(g, s) \mid g \,\mathrm{I}\, b \text{ or } g \,\mathrm{I}\, a\}$. Furthermore we denote the set of extents of $\mathbb{K}_{00}$ by $\mathrm{Ext}(\mathbb{K}_{00})$. We also set

$$\begin{aligned} \mathcal{H}(a) :=& \{A \cap a' \mid A \in \mathrm{Ext}(\mathbb{K}_{00}) \text{ and } A \cap a' \notin \mathrm{Ext}(\mathbb{K}_{00})\}\,, \\ \mathcal{H}(b) :=& \{A \cap b' \mid A \in \mathrm{Ext}(\mathbb{K}_{00}) \text{ and } A \cap b' \notin \mathrm{Ext}(\mathbb{K}_{00})\}\,, \\ \mathcal{H}(a \cup b) :=& \{A \cap (a' \cup b') \mid A \in \mathrm{Ext}(\mathbb{K}_{00}) \text{ and } A \cap (a' \cup b') \notin \mathrm{Ext}(\mathbb{K}_{00})\}\,, \\ \mathcal{H}(a \cap b) :=& \{A \cap (a' \cap b') \mid A \in \mathrm{Ext}(\mathbb{K}_{00}) \text{ and } A \cap (a' \cap b') \notin \mathrm{Ext}(\mathbb{K}_{00})\}\,. \end{aligned}$$

We will often write $h(x)$ for $|\mathcal{H}(x)|$, for any $x \in \{a, b, a \cap b, a \cup b\}$. Before we start the construction, let us recall the following result partly proved in [8]:

Theorem 1. *Let $\mathbb{K} := (G, M, \mathrm{I})$ be an attribute reduced context with $|G| \geq 3$ and $|M| > 3$. Let a and b be two attributes such that their existential generalization $s = a \cup b$ increases the size of the concept lattice. Then*

a) $|\mathfrak{B}(\mathbb{K})| = |\mathfrak{B}(\mathbb{K}_{00})| + |\mathcal{H}(a, b)|$, with $|\mathcal{H}(a, b)| = |\mathcal{H}(a) \cup \mathcal{H}(b) \cup \mathcal{H}(a \cap b)|$.
b) The increase is $|\mathcal{H}(a \cup b)| - |\mathcal{H}(a, b)| \leq 2^{|a'|+|b'|} - 2^{|a'|} - 2^{|b'|} + 1$.

Proof. Let $\mathbb{K} := (G, M, \mathrm{I})$ be such context and a, b two attributes of $\mathbb{K}$. One proceeds to the ∃-generalization of attributes a and b.

a) We set $\mathbb{K}^a = (G, M \setminus \{b\}, \mathrm{I})$. It holds:

$$|\mathfrak{B}(\mathbb{K})| = |\mathfrak{B}(\mathbb{K}^a)| + h^*(b) = |\mathfrak{B}(\mathbb{K}_{00})| + h(a) + h^*(b)$$

where $h^*(b) = |\{B \cap b';\ B \in \text{Ext}(\mathbb{K}^a),\ B \cap b' \notin \text{Ext}(\mathbb{K}^a)\}|$. Our aim is to express $h^*(b)$ as a function of $h(b)$ and $h(a \cap b)$. According to [8], $\text{Ext}(\mathbb{K}^a) = \text{Ext}(\mathbb{K}_{00}) \cup \mathcal{H}(a)$. Hence,

$$\begin{aligned}\mathcal{H}^*(b) &= \{B \cap b' \mid B \in \text{Ext}(\mathbb{K}^a), B \cap b' \notin \text{Ext}(\mathbb{K}^a)\}\\ &= \{B \cap b' \mid B \in \text{Ext}(\mathbb{K}_{00}) \text{ and } B \cap b' \notin \text{Ext}(\mathbb{K}^a)\}\\ &\quad \cup \{B \cap b' \mid B \in \mathcal{H}(a) \text{ and } B \cap b' \notin \text{Ext}(\mathbb{K}^a)\}\end{aligned}$$

Replacing $\text{Ext}(\mathbb{K}^a)$ by $\text{Ext}(\mathbb{K}_{00}) \cup \mathcal{H}(a)$, we get

$$\{B \cap b' \mid B \in \text{Ext}(\mathbb{K}_{00}) \text{ and } B \cap b' \notin \text{Ext}(\mathbb{K}^a)\} = \mathcal{H}(b) \setminus \mathcal{H}(a) \quad \text{and}$$

$$\{B \cap b' \mid B \in \mathcal{H}(a) \text{ and } B \cap b' \notin \text{Ext}(\mathbb{K}^a)\} = \mathcal{H}(a \cap b) \setminus (\mathcal{H}(b) \cup \mathcal{H}(a)).$$

$$\begin{aligned}\text{Thus,}\quad h^*(b) &= h(b) + h(a \cap b) - |\mathcal{H}(a) \cap \mathcal{H}(b)| + |\mathcal{H}(a \cap b) \cap \mathcal{H}(a) \cap \mathcal{H}(b)|\\ &\quad - |\mathcal{H}(a \cap b) \cap \mathcal{H}(a)| - |\mathcal{H}(a \cap b) \cap \mathcal{H}(b)|.\end{aligned}$$

Hence,

$$\begin{aligned}|\mathfrak{B}(\mathbb{K})| &= |\mathfrak{B}(\mathbb{K}_{00})| + |\mathcal{H}(a)| + |\mathcal{H}(b)| + |\mathcal{H}(a \cap b)| + |\mathcal{H}(a \cap b) \cap \mathcal{H}(a) \cap \mathcal{H}(b)|\\ &\quad - |\mathcal{H}(a) \cap \mathcal{H}(b)| - |\mathcal{H}(a \cap b) \cap \mathcal{H}(a)| - |\mathcal{H}(a \cap b) \cap \mathcal{H}(b)|\\ &= |\mathfrak{B}(\mathbb{K}_{00})| + |\mathcal{H}(a) \cup \mathcal{H}(b) \cup \mathcal{H}(a \cap b)|.\end{aligned}$$

b) Although b) was proved in [8], we can now get it from a). To maximize the increase $a' \cap b'$ should be $\emptyset$; i.e. $|\mathcal{H}(a \cap b)| \in \{0, 1\}$.

- If $|\mathcal{H}(a \cap b)| = 0$, then

$$\begin{aligned}|\mathfrak{B}(\mathbb{K})| &= |\mathfrak{B}(\mathbb{K}_{00})| + |\mathcal{H}(a) \cup \mathcal{H}(b) \cup \mathcal{H}(a \cap b)|\\ &= |\mathfrak{B}(\mathbb{K}_{00})| + |\mathcal{H}(a)| + |\mathcal{H}(b)|.\end{aligned}$$

- If $|\mathcal{H}(a \cap b)| = 1$, then we consider two subcases:
  - The only element of $\mathcal{H}(a \cap b)$ is not in $\mathcal{H}(a) \cup \mathcal{H}(b)$. Then,

$$\begin{aligned}|\mathcal{H}(a) \cap \mathcal{H}(b)| &= |\mathcal{H}(a \cap b) \cap \mathcal{H}(a) \cap \mathcal{H}(b)|\\ &= |\mathcal{H}(a \cap b) \cap \mathcal{H}(a)| = |\mathcal{H}(a \cap b) \cap \mathcal{H}(b)| = 0\end{aligned}$$

  and $|\mathfrak{B}(\mathbb{K})| = |\mathfrak{B}(\mathbb{K}_{00})| + |\mathcal{H}(a)| + |\mathcal{H}(b)| + |\mathcal{H}(a \cap b)|$.
  - The only element of $\mathcal{H}(a \cap b)$ is either in $\mathcal{H}(a)$ or $\mathcal{H}(b)$. Then

$$|\mathcal{H}(a \cap b)| + |\mathcal{H}(a \cap b) \cap \mathcal{H}(a) \cap \mathcal{H}(b)| - |\mathcal{H}(a \cap b) \cap \mathcal{H}(a)| - |\mathcal{H}(a \cap b) \cap \mathcal{H}(b)|$$

  is equal to zero and $|\mathcal{H}(a) \cap \mathcal{H}(b)| \in \{0, 1\}$. Thus

$$|\mathfrak{B}(\mathbb{K})| = |\mathfrak{B}(\mathbb{K}_{00})| + |\mathcal{H}(a)| + |\mathcal{H}(b)| + 1 - |\mathcal{H}(a) \cap \mathcal{H}(b)|.$$

In all these subcases, considering that $|\mathfrak{B}(\mathbb{K}_{0s})| = |\mathfrak{B}(\mathbb{K}_{00})| + |\mathcal{H}(a \cup b)|$, the increase after the generalization is

$$\begin{aligned} |\mathfrak{B}(\mathbb{K}_{0s})| - |\mathfrak{B}(\mathbb{K})| &= |\mathcal{H}(a \cup b)| - |\mathcal{H}(a, b)| \\ &\leq 2^{|a'|+|b'|} - 2^{|a'|} - 2^{|b'|} + (d_1 + d_2 - d_0) \\ &\leq 2^{|a'|+|b'|} - 2^{|a'|} - 2^{|b'|} + 1, \text{ since } d_1 + d_2 - d_0 \leq 0, \end{aligned}$$

with $d_1 = |\{A \subseteq a' \mid A \in \mathrm{Ext}(\mathbb{K}_{00})\}|$, $d_2 = |\{A \subseteq b' \mid A \in \mathrm{Ext}(\mathbb{K}_{00})\}|$ and $d_0 = |\{A \subseteq a' \cup b' \mid A \in \mathrm{Ext}(\mathbb{K}_{00})\}|$. □

Now, we define the following gain function:

$$\begin{aligned} \psi : M \times M &\longrightarrow \mathbb{Z} \\ (a, b) &\longmapsto \psi(a, b) = |\mathcal{H}(a \cup b)| - |\mathcal{H}(a, b)| \end{aligned}$$

Note that $\mathcal{H}(a \cup b) = \mathcal{H}(b \cup a)$, and $\mathcal{H}(a, b) = \mathcal{H}(b, a)$ because the order of adding the attributes a and b does not matter. Therefore $\psi(a, b) = \psi(b, a)$. By definition, $\psi(a, a) = 0$. Further, we define the map δ as followed:

$$\begin{aligned} \delta : M \times M &\longrightarrow \mathbb{R} \\ (a, b) &\longmapsto \begin{cases} 1 & \text{if } \ \psi(a, b) \leq 0 \\ 0 & \text{else} \end{cases} \end{aligned}$$

Since $\mathbb{K}$ is a finite context, there is a pair of attributes a_0, b_0 in M such that

$$|a_0'| + |b_0'| = \max_{a,b \in M} (|a'| + |b'|).$$

We set $n_0 = 2^{|a_0'|+|b_0'|} - 2^{|a_0'|} - 2^{|b_0'|} + 1$. Then $n_0 \geq 2^{|a'|+|b'|} - 2^{|a'|} - 2^{|b'|} + 1$ for all pairs $\{a, b\} \subseteq M$. With the function δ, we construct the following map:

$$\begin{aligned} S_{\text{gen}} : M \times M &\longrightarrow \mathbb{R} \\ (a, b) &\longmapsto S_{\text{gen}}(a, b) = \tfrac{1+\delta(a,b)}{2} - \tfrac{|\psi(a,b)|}{2n_0} \end{aligned}$$

where $|\psi(a, b)|$ is the absolute value of $\psi(a, b)$. That leads to the following results.

Proposition 1. *Let (G, M, I) be a reduced context with $|G| \geq 3$ and $|M| > 3$. Then S_{gen} is a normalized similarity measure on M.*

Proof. Let a, b two attributes of (G, M, I). Since $|\psi(a, b)| \leq n_0$ we can easily check that $0 \leq S_{\text{gen}}(a, b) = S_{\text{gen}}(b, a) \leq S_{\text{gen}}(a, a) = 1$ holds. □

S_{gen} also has the following properties:

Proposition 2. *Let (G, M, I) be a reduced context with $|G| \geq 3$ and $|M| > 3$. Let $a, b, c, d \in M$. It holds:*

a) $S_{gen}(a, b) \geq \frac{1}{2}$ if and only if $\psi(a, b) \leq 0$.

b) If $\psi(a,b) \leq 0 < \psi(d,c)$ *then* $S_{gen}(d,c) < S_{gen}(a,b)$.
c) If $0 < \psi(a,b) \leq \psi(d,c)$ *then* $S_{gen}(d,c) \leq S_{gen}(a,b)$.
d) If $\psi(a,b) \leq \psi(d,c) \leq 0$ *then* $S_{gen}(a,b) \leq S_{gen}(d,c)$.

Proof. Let $\mathbb{K} = (G, M, I)$ be such a context and $a, b, c, d \in M$.

a) If $\psi(a,b) \leq 0$ then $\delta(a,b) = 1$ and

$$S_{\text{gen}}(a,b) = \frac{1+\delta(a,b)}{2} - \frac{|\psi(a,b)|}{2n_0} = \frac{1}{2}\left(2 + \frac{\psi(a,b)}{n_0}\right) \geq \frac{1}{2}.$$

Now, $S_{\text{gen}}(a,b) \geq \frac{1}{2}$ implies $\frac{1+\delta(a,b)}{2} - \frac{|\psi(a,b)|}{2n_0} \geq \frac{1}{2}$ and $|\psi(a,b)| \leq n_0\delta(a,b)$. If $\delta(a,b) = 0$ then $|\psi(a,b)| = 0$. If $\delta(a,b) = 1$ then $\psi(a,b) \leq 0$ by definition of δ. Hence, $S_{\text{gen}}(a,b) \geq \frac{1}{2}$ if and only if $\psi(a,b) \leq 0$.

b) If $\psi(a,b) \leq 0 < \psi(d,c)$ then $S_{\text{gen}}(d,c) < \frac{1}{2} \leq S_{\text{gen}}(a,b)$.

c) If $0 < \psi(a,b) \leq \psi(d,c)$ then $\delta(a,b) = \delta(d,c) = 0$, and

$$S_{\text{gen}}(d,c) = \frac{1}{2} - \frac{\psi(d,c)}{2n_0} \leq \frac{1}{2} - \frac{\psi(a,b)}{2n_0} = S_{\text{gen}}(a,b).$$

d) If $\psi(a,b) \leq \psi(d,c) \leq 0$ then $\delta(a,b) = \delta(d,c) = 1$, and

$$S_{\text{gen}}(a,b) = 1 + \frac{\psi(a,b)}{2n_0} \leq 1 + \frac{\psi(d,c)}{2n_0} = S_{\text{gen}}(d,c).$$

□

Proposition 3. *Let* (G, M, I) *be a reduced context and* $a, b \in M$. *The following assertions are equivalent:*

(i) $\delta(a,b) = 1$.
(ii) $\psi(a,b) \leq 0$.
(iii) $S_{gen}(a,b) \geq \frac{1}{2}$.
(iv) *A* $\exists$*-generalization of* a *and* b *does not increase the size of the concept lattice.*

Proof. (i) $\Longleftrightarrow$ (ii) follows from the definition of δ. (ii) $\Longleftrightarrow$ (iii) is Proposition 2 a). (ii) $\Longleftrightarrow$ (iv) follows from the fact that $\psi(a,b) = |\mathcal{H}(a \cup b)| - |\mathcal{H}(a,b)|$ is actually the difference $|\mathfrak{B}(G, M \cup \{s\} \setminus \{a,b\}, I)| - |\mathfrak{B}(G,M,I)|$ between the number of concepts before and after generalizing a, b to s with $s' = a' \cup b'$.

Therefore, generalizing two attributes a, b in a reduced context (G, M, I) increases the size of the lattice if and only if $S_{\text{gen}}(a,b) < \frac{1}{2}$. The threshold $\frac{1}{2}$ is just a consequence of the way S_{gen} has been defined.

To test our results we have designed a naive algorithm (see Algorithm 1) that computes S_{gen} on all pairs of attributes a, b of $\mathbb{K}$. If the set of attributes M is considered as a vector, then for any attribute $a \in M$, we set T(a) the set of all attributes coming before a in M. The complexity of our algorithm is given by

$$\sum_{a \in M} (1 + \sum_{b \in M \setminus T(a)} ((q(a,b) + 4)[4(q(a,b) + 1) + 4] + 3),$$

which is equal to

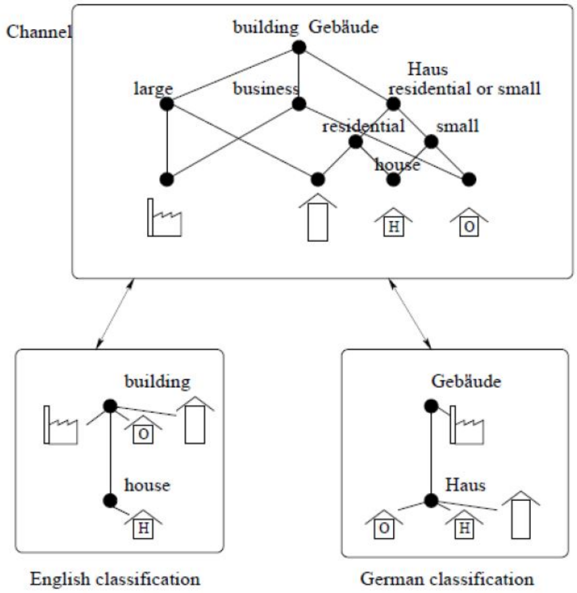
$$|M| + \sum_{a\in M}\sum_{b\in M\setminus T(a)} (4q^2(a,b)+24q(a,b)+35), \qquad \text{with } q(a,b) = |\,\mathrm{Ext}(\mathbb{K}_{00})|.$$

Algorithm 1: Computing a similarity measure

```
Data: An attribute reduced context $(G, M, I)$
Result: $\psi$ and $S_{\text{gen}}$ on $M \times M$
Choose $x, y$ in $M$, $x \neq y$ with $|x'| + |y'|$ maximal;
$n_0 \leftarrow 2^{|x'|+|y'|} - 2^{|x'|} - 2^{|y'|} + 1$;
$T \leftarrow \emptyset$;
foreach $a$ in $M$ do
    $T \leftarrow T \cup \{a\}$;
    foreach $b$ in $M \setminus T$ do
        $\mathrm{Ext}_0 \leftarrow \mathrm{Ext}(G, M \setminus \{a, b\}, I)$;
        foreach $x$ in $\{a, b, a \cup b, a \cap b\}$ do $\mathcal{H}(x) \leftarrow \emptyset$;
        foreach $A$ in $\mathrm{Ext}_0$ do
            foreach $x$ in $\{a, b, a \cup b, a \cap b\}$ do
                if $A \cap x' \notin \mathrm{Ext}_0$ then $\mathcal{H}(x) \leftarrow \mathcal{H}(x) \cup \{A \cap x'\}$;
            end
        end
    end
    $\psi(a, b) \leftarrow |\mathcal{H}(a \cup b)| - |\mathcal{H}(a) \cup \mathcal{H}(b) \cup \mathcal{H}(a \cap b)|$;    $\psi(b, a) \leftarrow \psi(a, b)$;
    if $\psi(a, b) \leq 0$ then
        $\delta(a, b) \leftarrow 1$
    else
        $\delta(a, b) \leftarrow 0$
    end
    $S_{\text{gen}(a,b)} \leftarrow \dfrac{1 + \delta(a, b)}{2} - \dfrac{|\psi(a, b)|}{2n_0}$
end
```

4 An Example from Lexicographic Data

Formal Concept Analysis has been applied to compare lexical databases. In [11] Uta Priss proposes an example in where the information channel is "*building*". With respect to this, the main difference between English and German is that in English, the word "house" only refers to small residential buildings whereas in German even small office buildings and large residential buildings can be called "Haus", and only factories would normally not be called "Haus". Moreover, "building" in English refers to either a factory, an office or even a big residential house. But only a factory can be called "Gebäude" in German. She presented in the figure below the information channel of the word "building" in the sense of Barwise and Seligman [2] in both English and German.



With the above information channel we can construct a formal context as follows: The objects are different kinds of buildings: small house ("h"), office ("o"), factory ("f") and large residential house ("l"). The attributes are different names of these objects in both languages: English and German. These are "building", "house", "Haus", "Gebäude", "large building" (short: "large"), "business building" (short: "business"), "residential house" (short: "residential"), and "small house" (short: "small"). Thus $G = \{$h, o, f, l$\}$ and $M = \{$"building", "house", "Haus", "Gebäude", "large", "business", "residential", "small"$\}$. In the following, a set of objects will be denoted as a concatenation of those objects. For example we will write ho or oh for the set $\{$h, o$\}$. The English and German classifications of the word "building" are then presented in the following formal context:

	building	house	Haus	Gebäude	large	business	residential	small
factory	×			×	×	×		
office	×		×			×		×
house		×	×				×	×
large	×		×		×		×	

For this formal context, $n_0 = 2^{3+3} - 2^3 - 2^3 + 1 = 49$. Let consider the attributes $a :=$ house and $b :=$ Gebäude. Then $a' \cup b' = \{f, h\}$ and $a' \cap b' = \emptyset$. We have

$$\mathrm{Ext}(\mathbb{K}_{00}) = \{fohl, fol, ohl, fo, fl, ol, oh, hl, f, o, h, l, \emptyset\}, \text{ and}$$

$\mathcal{H}(a) = \mathcal{H}(b) = \mathcal{H}(a \cap b) = \emptyset$ and $\mathcal{H}(a \cup b) = \{fohl\}$. Therefore, $\psi(a, b) = 1$ and $S_{\text{gen}}(a, b) = \frac{1}{2} - \frac{1}{98} \approx 0.49$. Using our algorithm, we compute $\psi(a, b)$ and

$S_{\text{gen}}(a, b)$ for all pairs $a, b \in M$. The table below show $\psi(a, b)$ below the diagonal, and $S_{\text{gen}}(a, b)$ on the rest.

	building	house	Haus	Gebäude	large	business	residential	small
building	1.00	0.98	0.97	1.00	0.99	0.98	0.97	0.97
house	−2	1.00	1.00	0.49	0.49	0.49	1.00	1.00
Haus	−3	0	1.00	0.98	0.97	0.97	0.99	0.99
Gebäude	0	1	−2	1.00	1.00	1.00	0.49	0.49
large	−1	1	−3	0	1.00	0.98	0.49	0.97
business	−2	1	−3	0	−2	1.00	0.98	0.49
residential	−3	0	−1	1	1	−2	1.00	0.98
small	−3	0	−1	1	−3	1	−2	1.00

From the above table, the attributes ”house” and ”Gebäude” are less similar. It reflects the fact that these words ”Gebäude” (in German) and ”house” (in English) do not have the same meaning. It is also the case for the attributes ”house” and ”business buildings” as well as ”Gebäude” and ”residential building”. Hence, putting together each of the above pairs of attributes will increase the size of the lattice. On the contrary, the attributes ”large” and ”Haus”, ”building” and ”Haus” are more similar through S_{gen}. It is because the word ”Haus” which designates a house, a business office or simply large building in German, often coincides with the words ”building” or ”large building” in English. For these pairs, the existential generalization will not increase the size of the lattice.

5 Conclusion

We have constructed a similarity measure compatible with the change in the size of the lattice after a generalization of a pair of attributes in a formal context. That measure should send a warning when grouping two attributes. Also, it enables us to characterize contexts where generalizing two attributes increases the size of the concept lattice. Our next step is to look at the implication between generalized attributes. We suspect that the number of implications decreases if the number of concepts increases.

References

1. Alqadah, F., Bhatnagar, R.: Similarity Measures in Formal Concept Analysis. AMAI – Springer (2009)
2. Barwise, J., Seligman, J.: Information Flow: the logic of distributed systems. Cambridge University Press (1997)
3. Besson, J., Pensa, R. G., Robardet, C., Boulicaut, J.: Constraint-Based Mining of Fault-Tolerant Patterns from Boolean Data. KDID 55–71 (2005)
4. Dice, L. R.: Measures of the Amount of Ecologic Association Between Species. esa. Promoting the Science of Ecology Vol. 26, No 3, 297–302 (1945)

5. Domenach, F.: Similarity Measures of concept lattices In: Lausen B., Krolak-Schwerdt S., Böhmer M. (eds) Data Science, Learning by Latent Structures, and Knowledge Discovery. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg
6. Kwuida, L.: On concept lattice approximation. In Osamu Watanabe (Ed.) Foundations of Theoretical Computer Science: For New Computational View. RIMS No. 1599, Proceedings of the LA Symposium (2008), January 28–30, Kyoto, Japan, 42–50 (2008)
7. Kwuida, L., Missaoui, R., Balamane A., Vaillancourt, J.: Generalized pattern extraction from concept lattices. AMAI – Springer 151–168 (2014)
8. Kwuida, L., Kuitché, R. S., Temgoua, E. R. A.: On the Size of ∃-Generalized Concepts. ArXiv:1709.08060.
9. Ganter, B., Wille, R.: Formal Concept Analysis. Mathematical Foundations. Springer (1999)
10. Jaccard, P.: Nouvelle recherche sur la distribution florale. Bulletin de la Société Vaudoise des Sciences Naturelles (1908)
11. Priss, U.: Linguistic Applications of Formal Concept Analysis. Formal Concept Analysis, 149–160 (2005)
12. Rogers, D. J., Tanimoto, T. T.: A Computer Program for classifying plants. Springer (1960)
13. Sneath, P. H. A.: The Application of Computers to Taxonomy J. gen Microbiol. 17, 201–226 (1957)
14. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing iceberg concept lattices with Titanic. Data Knowl. Eng., 42(2): 189–222 (2002)
15. Ventos, V., Soldano, H., Lamadon, T.: Alpha Galois Lattices. ICDM, 555–558 (2004)
16. Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In I. Rival (Ed.) Ordered Sets. Reidel, 445–470 (1982)
17. Wille, R.: Lattices in data analysis: how to draw them with a computer. In I. Rival (Ed.) Algorithms and Order. Kluwer, 33–58 (1989)
18. Wille, R.: Tensorial decomposition of concept lattices. Order 2, 81–95 (1985)
19. Wille, R.: Subdirect product construction of concept lattices. Discrete Mathematics 63, 305–313 (1987)