Final ELIAS exchange grant Scientific Report

Francisco J. Valverde Albacete

Abstract

This report of activity concerns my ELIAS exchange grant from 07/15/2013 to 08/18/2013 at the i-Kernels group led by Prof. A. Moschitti in the Department of Information Engineering and Computer Science (DISI), at University of Trento, Italy (UNITN).

1 Purpose of the visit

The following specific scientific goals were explicitly addressed by this exchange grant:

- In general, a study of the feasibility of the evaluation of Formal Concept Analysis (FCA) by information-theoretic means, and
- In particular, the assessment of FCA-enabled distributional models of Semantics by information-theoretic means.

However, since the funded period for the grant was halved, I have been obliged to limit myself mostly to the first one, for the time being. For the second avenue of research, work has only began on a proof of concept.

2 Description of work

This report concerns the development of an information-theoretic assessment model for a new kind of semantic models being developed in the framework of Project LiMoSINe, *Formal Concept Analysis (FCA)-enhanced distributional semantic models.*

Distributional models of language have reached a point where it is feasible to parameterise them around a number of design decisions [11]. UNED is experimenting with the enhancing of such models by incorporating Formal Concept Analysis-based capabilities [4]. The envisioned extension deals with the structuring of the dimensions of the distributional model to resemble the lexical dimensions of a lexical hierarchy, hence modelling ad-hoc lexical resources that lend explicability to distributional model features.

For this purpose it is necessary that we use graded incidences where the degree of incidence between objects and their attributes ranges in a continuous interval. However, not every range of measurements generate FCA-like theories for the analysis of data.

The author of this grant proposal has co-developed an extension of FCA where objects and attributes are related by degrees of incidence [20–22, 24] and a framework for evaluating multiclass classification in information-theoretic terms[23, 25]. It is our intuition that these two tools can be combined to provide a measure to assess the quality of a $\mathcal{K}$-FCA-induced concept lattices in semantic-modelling tasks.

2.1 Formal Concept Analysis to Enhance Distributional Models of Language

Formal Concept Analysis (FCA) is an unsupervised learning technique that generates concept lattices from binary matrices (incidence relations) between objects and attributes whereby their relation can be visualized or explored. The result of the FCA building process is a *concept lattice*, that is, a complete lattice of *formal concepts*, pairs of a set of objects and all the attributes (features) that can be *predicated* of them.

The mathematical bases of this induction process are well understood and robust, hence attractive for the induction of hierarchy-like structures (tree-like hierarchies are trivially embedded in complete lattices). Therefore, FCA as an exploratory data technique that has been used extensively in Linguistic Modelling [14], Knowledge Processing [13] or Information Retrieval [26], to cite but a few domains.

One could then use similarity in concept lattices [3] as modelling lexical similarity, and profit from the existence of object and attribute topologies in the induced lattice [6]. Since FCA induces new proximities and relations between objects (dually attributes) it would be possible to deduce relations unknown/unobserved in the data.

This process is reminiscent of Singular Value Decompositionand Latent Semantic Analysis [22] , but the dimensions of analysis are related to the original ones in the model. Therefore, to better capture linguistic phenomena, the original dimensions of the matrix must be linguistically-motivated. The technology of tree kernels used at the Machine Learning and NLP group at U. Trento has proven its mettle in describing linguistic phenomena for standard text categorization tasks [1], and we expect it to excel in capturing phenomena for FCA-amenable features.

2.2 Extending FCA to model linguistic, learnable phenomena

From the perspective of Machine Learning, FCA is a subtask of multilabel classification [17, 18], specifically, the exploratory analysis of data/classification results. It has many points of contact with hierachical clustering and, more specifically, biclustering. Under this guise, FCA is also known as *conceptual clustering.*

But it has proven difficult to develop it as an objective function optimization-driven unsupervised or supervised learning task [but see 19, as applied to Information Retrieval]. This is, in part, caused by the absence of graded assessment measures on concept lattices: in standard FCA a concept lattice either represents is incidence relation or not, suggesting a discrete measure.

$\mathcal{K}$-FCA is an extension of FCA where incidences take value in a special type of range, specifically an idempotent semifield (logprobabilistic costs, for intance). Matrices with values in such algebras are subjected to an exploratory process that generates concepts with a certain *degree of existence*. The issue of this process is a sequence of concept lattices that can be related to such degrees of existence. Such technique has been used to explore confusion matrices of multiclass classifiers [12] as well as gene expression microarray data [7].

2.3 Information-theoretic methods to evaluate Machine Learning tasks

In previous work, we have co-developed a set of techniques to evaluate multiclass classifications tasks [23] based information-theoretic methods, namely the (De Finetti) Entropy Triangle (ET) and the Entropy-Modified Accuracy (EMA) and the Normalized Information Transfer (NIT) factor.

EMA is a measure of a classifier's performance in a classification task while the NIT factor is a measure of how efficient is the learning process of the classifier [25]. Furthermore, the ET is a visualizing tool for system comparison that takes into consideration both previous measures (indirectly) as well as the difficulty of the task (explicitly) [23]. Furthermore, it is also possible to analyse the performance of classifiers on a per-class basis by means of $\mathcal{K}$-FCA exploration of their confusion matrices [12].

Note that such measures can already be used as alternative performance indices for tasks that use distributional models, like detecting synonymy [11, §6]. However, such measures cannot be used *yet* for the analysis of multilabel classification, i.e. for the performance analysis of those matrices generating concept lattices in either binary or graded form. And yet, avowedly [2, 5], concept lattices improve on other tasks to evaluate distributional models, like sense ranking or single-word priming.

In this work we have undertaken **to extend the information-theoretic measures and visualization technique above to the evaluation of multilabelling tasks**, which amounts to the evaluation of certain concept lattices arising from them. This will impact on all tasks interpretable as multilabelling such as text categorization, market basket analysis, etc.

2.4 FCA for the evaluation of multi-labelling tasks

The technique to evaluate classification tasks by information-theoretic means is crucially based in the availability of probability densities over a sigma algebra [23]. In the finite case, sigma algebras reduce to Boolean lattices, but defining probability measures or densities on generic lattices is a long-sought goal as yet unattained. In fact relaxations of the measure concept have been attained on geometric or merely distributed lattices, but not on general lattices [9].

FCA provides a re-foundational basis for lattice theory in terms of concrete lattices: every lattice is the concept lattice of some data context and every data concept has an associated context lattice. In fact, the fundamental theorem of FCA provides some guidance as to how to define functions over lattices by considering the sets of join- and meet-irreducibles, unlike standard measures over boolean lattices which only use join-irreducibles (atomic events in the finite sigma-field of sets).

2.5 Extending measure theory to Concept Lattices

Generalising the definition of measure of probability over a sigma algebra, or boolean lattice, to other valuation functions on more general types of lattices is a long-sought goal of a number of disciplines [27], but so far elusive. In the framework of the Dempster-Schafer Theory of evidence some advance has been made by considering belief functions to take values in *semifields*, for instance ranking theory [16]. Semifields are abstractions of both the concepts of semirings and fields, some of them lacking most of the properties normally handled when calculating in rings or fields. On the other hand, the boolean semiring, extensively used in Computer Science and Combinatorics, is a good prototype for such semirings, lacking additive and multiplicative inverses, but with other interesting properties. In fact, it seems that FCA is nothing but a different take on the (almost inexistent) spectral theory over certain semirings [22].

In this paper we propose to explore this issue from the point of view of an extension to lattice theory by considering valuations over concept lattices over semirings.

2.5.1 Classical measure theory definitions

Let G be a non-empty set and $\Gamma \subseteq 2^G$ be a non-empty class of subsets of G. Recall that a (classical) measure is a function $\mu : \Gamma \to [0, \infty]$ assigning to every subset a nonnegative number $X \mapsto \mu(X)$. Then $X \in \Gamma$ is a null set iff $\mu(X) = 0$.

The natural extension of measures are valuations over lattices: Let $\mathcal{L} = \langle L, \vee, \wedge, \bot_L, \top_L \rangle$ be a non-empty, finite lattice. An *evaluation on* $\mathcal{L}$ is a map $r : L \to \mathbb{R}$. We say that an evaluation r on $\mathcal{L}$ is:

1. *isotone*, if $x < y \Rightarrow r(x) \leq r(y)$, and *strictly isotone*, if $x < y \Rightarrow r(x) < r(y)$.
2. *antitone*, if $x < y \Rightarrow r(x) \geq r(y)$, and *strictly antitone*, if $x < y \Rightarrow r(x) > r(y)$.
3. *grounded*[1] when $r(\bot_L) = 0$, and *normalized* when $r(\top_L) = 1$.
4. *submodular*, iff $r(x \vee x) + r(x \wedge y) \leq r(x) + r(y)$.
5. *supermodular*, iff $r(x \vee x) + r(x \wedge y) \geq r(x) + r(y)$.
6. *modular*, iff it is both submodular and supermodular.

A *valuation* on $\mathcal{L}$ is an isotone modular evaluation on $\mathcal{L}$. See [10] for a review of the importance of valuations in quantitative analysis. Note that [15] considers antitone modular evaluations, calling them *entropies*. This dovetails into the concept of Galois connections made concrete by FCA.

[10] propose a strictly isotonic, submodular evaluation:

$$t(a) = \begin{cases} 0 & a = \top_L \\ t(a^*) - 1 & a \in \mathcal{M}(L), a < a^* \\ \min\{t(b) + t(c) - t(b \vee c) \| a < b, a < c\} & \text{otherwise} \end{cases} \quad (1)$$

Note that this valuation seems to be defined in $\mathbb{N}$ rather than $\mathbb{R}$. Clearly, their intent was to relate a closure and a kernel system through it. That is, to give an overall valuation for a concept lattice in terms either of the closure system of extents or the kernel (dual closure) system of intents. However, this was not achieved in [10] or elsewhere, to the extent of our knowledge.

[1] [10] use *normalized*, but this is mostly used elsewhere in the same sense we use it here.

2.5.2 Valuations on a concept lattice

We introduce now a generalization of valuations where the range is any semiring. Let $\mathcal{S} = \langle S, \oplus, \otimes, \perp, e \rangle$ be a semiring and $\mathcal{L}$ a non-empty, finite lattice. A $\mathcal{S}$-evaluation on $\mathcal{L}$ is a map $r : L \to S$. We say that a $\mathcal{S}$-evaluation r on $\mathcal{L}$ is:

1. *grounded* when $r(\perp_L) = \perp_S$ and *normalized* when $r(\top_L) = e_S$.
2. *submodular*, iff $r(x \vee x) \otimes r(x \wedge y) \leq r(x) \otimes r(y)$.
3. *supermodular*, iff $r(x \vee x) \otimes r(x \wedge y) \geq r(x) \otimes r(y)$.
4. *modular*, iff it is both submodular and supermodular.

A $\mathcal{S}$-valuation on $\mathcal{L}$ is an isotone, modular $\mathcal{S}$-evaluation on $\mathcal{L}$.

Note that we have given a multiplicative character to the modular law. Then traditional modularity can be interpreted as being defined in the max plus semiring.

It is easy to see that in this case (1) can actually be conceived as defined over the $\overline{\mathbb{R}}_{\min,+}$ idempotent semifields: actually the complete subsemifield with carrier set $[-\infty, 0]$ which is an incline. In order to generalize it further we are going to posit a mass function for meet irreducibles $m_L : \mathcal{M}(L) \to S$, so that

$$t^{\uparrow}(a) = \begin{cases} e & a = \top_L \\ t^{\uparrow}(a^*) \ \dot{/} \ m_L(a) & a \in \mathcal{M}(L), a < a^* \\ \sum^{\bullet}_{b,c \in \uparrow a \setminus \{a\}} (t^{\uparrow}(b) \,\dot{\otimes}\, t^{\uparrow}(c)) \ \dot{/} \ t^{\uparrow}(b \vee c) & \text{otherwise} \end{cases} \tag{2}$$

Now consider a mass function on join-irreducibles $m_L^{\delta} : \mathcal{J}(L) \to S$. In such case we can define a dual valuation as

$$t^{\downarrow}(b) = \begin{cases} e & b = \perp_L \\ t^{\downarrow}(b^*) \ \dot{/} \ m_L^{\delta}(b) & a \in \mathcal{J}(L), b > b^* \\ \sum^{\bullet}_{b,c \in \downarrow a \setminus \{a\}} (t^{\downarrow}(b) \,\dot{\otimes}\, t^{\downarrow}(c)) \ \dot{/} \ t^{\downarrow}(b \vee c) & \text{otherwise} \end{cases} \tag{3}$$

With such valuation we may define a valuation for a concept (a, b) in a concept lattice as:

$$v : L \to S \tag{4}$$

$$(a, b) \mapsto v((a, b)) = (k_1 \,\dot{\otimes}\, t^{\uparrow}(a)) \ \dot{/} (k_2 \,\dot{\otimes}\, t^{\downarrow}(b)) \tag{5}$$

Note that both $t^{\uparrow}$ and $t^{\downarrow}$ are strict isotone and supermodular, whence the behaviour of v can be adjusted by means of the constants k_1 and k_2 and depends heavily on the mass functions on join- and meet-irreducibles.

2.6 Tools to measure the performance of multilabel classification

This section summarizes the present state of affairs regarding measuring performance for multilabel classification by using entropic measures on multi-valued formal contexts.

The mass functions used in the previous section have some bearing in the notion of "uncertainty" in the lattice: The process of reducing a formal context defines a partition of the set of objects $G/\ker(\gamma)$ and attributes $M/\ker(\mu)$ [4]. If the resulting concept lattice were boolean, such partitions can be used to define probability mass functions on the atoms and coatoms of the lattice, respectively, corresponding to the meet and the join irreducibles.

If the partition of objects had a single block, the ability of the attribute set to distinguish between objects would be the worst possible, yielding a system with maximal *object uncertainty.* Contrariwise, if the partition had as many blocks as objects, the ability of the attribute set to distinguish objects would be the best possible, yielding a system with minimal object uncertainty. A dual discussion would lead to similar considerations for *attribute uncertainty.* Note that the uncertainty of such partitions is both related to the Hartley and the Shannon entropies [9]

With respect to our previous work on measures on joint distributions, these object and attribute uncertainty seem to capture the notion of *(marginal density) distance with respect to the uniform distribution* [23]. This lends credibility to the notion that there exists a similar "entropy triangle" operating on formal contexts as the one operating on joint distributions.

For instance, it is well known that a perfect coupling between objects and attributes is captured by diagonal incidences, equivalently, diamond lattices. Dually, the worst coupling between objects and attributes is expressed in formal contexts whose concept lattices are boolean (maximal uncertainty). Note the resemblance of these behaviours to random variable independence. The sum of the residual conditional entropies in our entropy triangle is the analogue of this behaviour. Transposed to the FCA context it actually models the "uncoupling" between objects and attributes (e.g., what cannot be discerned about one with the other, in either direction).

The third dimension that an "uncertainty triangle" would need is the analogue of the mutual information in a joint distribution. By analogy, this is what can be known about objects by their attributes and vice-versa. I have not been able to find a closed expression to this final quantity and this is the reason why this discussion has been kept qualitative.

2.7 FCA for modelling linguistic learnable phenomena

This is the avenue of research that has suffered most from the reduction in the initial requested duration for the grant. Nevertheless, since this is the point where the host institution was strongest we have decided to sketch the lines of future joint research rather than carry out the brunt of it.

Consequently, we have decided to use UNITN's expertise in tree kernels and UNED's expertise in FCA for the task of developing features for classification tasks using tree kernels and to apply it to Relation Extraction tasks.

The idea is for UNED's to encode gold standard relations as lattices and to obtain tree features both from the meet- and the join-semilattice that are included in the concept lattice, to supplement feature from other linguistic data available. The feature-choosing capabilities of tree-kernels will then be used to choose the more informative features. The expectation is that the combination of more concrete or abstract features—as provided by meet- or join-semilattices—will provide a richer set of informations to extract relations with.

At present we are building a proof-of-concept experiment on the ACE data [8].

3 Description of the main results

We have generalised the concept of a valuation on $\mathbb{R}_0^+$ over a lattice to coupled valuations on idempotent semifields over concept lattices (Section 2.5)

We have further explored the concept of an entropy triangle (Section 2.6), similar to previous work of ours, that captures the concept of uncertainty in Concept Lattices in several different ways:

- In the inherent partitioning entailed by the definition of concepts,
- In the particular structure of the concept lattice and how closely it resembles a diamond or a boolean lattice.

The quantitative description of this tool is left pending.

We have laid out the research avenue to obtain information from concept lattices that tree kernels can use (Section 2.7).

4 Future collaboration with host institution (if applicable)

At present, collaboration with DISI, UNITN centres around project LiMoSINe, where both the LSI at UNED and iKernels at UNITN are partners in the developing of a deep semantics distributional model. We plan to include several of the results of this research stay in that model.

The group of M. Baroni, at CiMEC, UNITN would also be very interested in developments around lattice representations of linguistic data. Indeed, I was invited to give a talk at CiMEC, UNITN about this particular on 07/18/2013.

Apart from other collaborations, the extraction of features from concept lattices dovetails nicely into the tree-kernel approach to classification co-developed and perfected at the i-kernels group in DISI, UNITN. Consequently, we have devised a joint line of research in using attributes extracted from concept lattices to help in Relation Extraction tasks. This is detailed in Section 2.7.

5 Projected publications resulting from the grant

1. A journal paper on the development of measures on concept lattices that takes into consideration the closure and the kernel systems, as instantiated on completed semifields. Possible venues are Journal of Fuzzy Sets and Systems and Information Sciences.
2. A journal paper with the mathematical development of the uncertainty measures on lattices that also allow to measure the performance of multilabel classification. Same venues as previous paper.
3. A conference paper on the feature selection from concept lattices for Relation Extraction is envisioned for the second line of work.

References

[1] Robertro Basili and Alessandro Moschitti. *Automatic Text Categorization: From Information Retrieval to Support Vector Learning.* ARACNE editrice, 2005.

[2] Claudio Carpineto and Giovanni Romano. *Concept Data Analysis.* Theory and Applications. John Wiley & Sons, Ltd, Chichester, UK, September 2005.

[3] Anna Formica. Concept similarity in Formal Concept Analysis: An information content approach. *Knowledge-Based Systems*, 21(1):80–87, February 2008.

[4] Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis: Mathematical Foundations.* Springer, Berlin, Heidelberg, 1999.

[5] R. Godin, M. Missaoui, and A. April. Experimental comparison of navigation in a Galois lattice with conventional information retrieval methods. *International Journal of Man-Machine Studies*, 38:747–767, 1993.

[6] Michel Gondran and Michel Minoux. *Graphs, Dioids and Semirings. New Models and Algorithms.* Operations Research Computer Science Interfaces series. Springer, 2008.

[7] J.M. González-Calabozo, Carmen Peláez-Moreno, and Francisco J. Valverde-Albacete. Gene expression array exploration using $\mathcal{K}$-Formal Concept Analysis. In P. Valtchev and R. Jäschke, editors, *Proceedings of the ICFCA11*, volume 6628 of *LNAI*, pages 119–134. Springer, May, 1–6 2011.

[8] B. Hachey, C. Grover, and T. Tobin. Datasets for generic relation extraction. *Natural Language Engineering*, 2011.

[9] George J Klir. *Uncertainty and information: foundations of generalized information theory.* Wiley-Interscience, 2006.

[10] Léonard Kwuida and Stefan E Schmidt. Valuations and closure operators on finite lattices. *Discrete Applied Mathematics*, 159(10):990–1001, 2011.

[11] Sebastian Padó and Mirella Lapata. Dependency-Based Construction of Semantic Space Models. *Journal of the ACM*, 33(2):161–199, June 2007.

[12] Carmen Peláez-Moreno, Ana I. García-Moral, and Francisco J. Valverde-Albacete. Analyzing phonetic confusions using Formal Concept Analysis. *Journal of the Acoustical Society of America*, 128(3):1377–1390, Sep 2010.

[13] Jonas Poelmans, Sergei O Kuznetsov, D Ignatov, and Guido Dedene. Formal Concept Analysis in Knowledge Processing: a Survey on Models and Techniquesn (to appear). *Expert Systems with Applications*, 2013.

[14] U Priss. Linguistic applications of formal concept analysis. In Bernhard Ganter et al., editors, *Formal Concept Analysis*, pages 149–160. Springer Berlin Heidelberg, 2005.

[15] Dan A Simovici. Betweenness, Metrics and Entropies in Lattices. In *ISMVL 2008. 38th International Symposium on Multiple Valued Logic, 2008.*, pages 26–31. IEEE, 2008.

[16] Wolfgang Spohn. *The Laws of Belief: Ranking Theory and Its Philosophical Applications.* Oxford, 2012.

[17] G Tsoumakas, I Katakis, and Ioannis Vlahavas. Mining multi-label data. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Data Mining and Knowledge Discovery, 2010.

[18] Grigoris Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and ...*, 2007.

[19] Francisco J. Valverde-Albacete. Combining soft and hard techniques for the analysis of batch retrieval tasks. In Enrique Herrera-Viedma, Gabriella Pasi, and Fabio Crestani, editors, *Soft Computing for Information Retrieval on the Web. Models and Applications*, volume 197 of *Studies in Fuzziness and Soft Computing*, pages 239–258. Springer, 2006. ISSN 1434-9922 (print edition). ISSN (electronic edition) 1860-0808, ISBN-10 3-540-31588-8, ISBN-13 978-3-540-31588-9.

[20] Francisco J. Valverde-Albacete and Carmen Peláez-Moreno. Further Galois connections between semimodules over idempotent semirings. In Jean Diatta and Peter Eklund, editors, *Proceedings of the 4th Conference on Concept Lattices and Applications (CLA 07)*, pages 199–212, Montpellier, October 2007.

[21] Francisco J. Valverde-Albacete and Carmen Peláez-Moreno. Galois connections between semimodules and applications in data mining. In S.O. Kusnetzov and S. Schmidt, editors, *Formal Concept Analysis. Proceedings of the 5th International Conference on Formal Concept Analysis, ICFCA 2007, Clermont-Ferrand, France*, volume 4390 of *LNAI*, pages 181–196. Springer, 2007.

[22] Francisco J Valverde-Albacete and Carmen Pelaez-Moreno. Spectral lattices of $(R_{max,+})$-formal contexts. *Formal Concept Analysis. Proceedings of the 6th International Conference, ICFCA 2008, Montreal, Canada, February 25-28, 2008.*, Lecture Notes in Computer Science(LCNS 4933):124–139, 2008.

[23] Francisco J. Valverde-Albacete and Carmen Peláez-Moreno. Two information-theoretic tools to assess the performance of multi-class classifiers. *Pattern Recognition Letters*, 31(12):1665–1671, 2010.

[24] Francisco J. Valverde-Albacete and Carmen Peláez-Moreno. Extending conceptualisation modes for generalised Formal Concept Analysis. *Information Sciences*, 181:1888–1909, May 2011.

[25] Francisco J. Valverde-Albacete and Carmen Peláez-Moreno. The normalized information transfer factor measures classification task performance or 100% accuracy considered harmful. *In preparation*, 2013.

[26] Francisco J. Valverde-Albacete and Carmen Peláez-Moreno. Systems vs. methods: an analysis of the affordances of formal concept analysis for information retrieval. In *Proceedings of the Workshop "Formal Concept Analysis meets Information Retrieval" (FCAIR)*, 2013.

[27] Zhenyuan Wang and George J Klir. *Generalized measure theory.* Springer Berlin Heidelberg, January 2008.