

Hyper-Rectangular Relation Decomposition and Dimensionality Reduction

F. Ferjani¹, A. Jaoua¹, S. Elloumi¹, and S. Ben Yahia²

¹ Qatar University, Qatar

² Faculté des Sciences de Tunis, Tunisia

Abstract. Among the different binary relation decompositions rectangular, pseudo-rectangular and Fringe based methods have been investigated these last years for labeling the main concepts of a document or a corpus. In this paper, we propose a new heuristic based on hyper rectangles for navigating and browsing in a corpus. A Hyper Rectangle is defined as the union of all possible rectangles in a binary relation involving an element of the domain or the range of given binary relation. From any binary relation \mathcal{R} , we can efficiently derive a maximum of $(n+m)$ hyper rectangles, where n (respectively m) is the cardinality of the domain (respectively the range) of \mathcal{R} . By applying a heuristic for a multilevel minimal coverage of relation \mathcal{R} , a browsing tree in the relation is generated for a meaningful navigation. We can also apply such decomposition to find different summarization levels in a document at a micro level. The same method is also applied for a macro level structuring of a corpus.

Keywords: Rectangular decomposition, hyper rectangle, fringe relation, corpus structuring, text summarization, feature extraction, text mining.

1 Introduction

During the last decades, the extensive use of computer-based corpora for a range of language studies has led to the exploration of the optimal ways in which texts within a large corpus are organized. In this way, it is so important to organize and structure the corpus and make later navigating and browsing become more easy, friendly and efficient. Structuring/Browsing is one of the most popular ways to gather and structure a corpus. The organization of the large document corpus is the problem we concern. The goals to tackle in this paper consisted to resolve the following problems. Firstly, it concerns the high dimensionality which bring memory and time complexity challenges to later clustering algorithms. Secondly, it concerns the ambiguity: synonyms and words with multiple meaning are very common. Thirdly, the structuring/browsing through different levels such that:

- Macro-level browsing deals with the global visual and logical structure of the corpus (e.g. categories, sub-categories, sections, fields, ...)
- Micro-level browsing is used for navigating through a textual document. For example, company information, location, ...

The micro and macro-level browsing require different methods for performing automatic navigation: Whereas macro-level browsing is mainly based on the pertinent information inside the corpus, micro-level browsing typically requires basic linguistic processing inside a sub-corpus or a document. Most of existing systems using conceptual analysis are NP-complete [1] and only able to analyze a small number of documents and/or web pages [2] [3]. In this paper, we propose a novel approach for text structuring that should only require a linear time in terms of the size of the binary context linking documents to indexing words or sentences inside a same document organized as Hyper Conceptual Rectangle (i.e., *hyperconcept* which is very appropriate for large data set and may overcome the previous approaches drawbacks). The paper is organized as the following: the section 2 includes some relational algebra, and formal concept analysis; the mathematical foundations used in this work. In section 3, we present how to build a generic tree of words through which user can browse easily to find the most pertinent documents in decreasing order of their importance. Finally, we concludes and points out avenues for future work concerning the heuristics for text mining and textual structuring.

2 Key Settings and HYPER RECTANGLE DEFINITION

Relational Algebra and Formal Concept Analysis may be considered as useful mathematical foundations that unified data and knowledge in information retrieval systems.

Binary Relations In the following, we recall some basic definitions from relational algebra [4]:

- A relation \mathcal{R} is a subset of the cartesian product of two sets \mathcal{X} and \mathcal{Y} .
- An element $(e, e') \in \mathcal{R}$ where e' denotes the image of e by \mathcal{R} .
- A binary relation Identity $\mathcal{I}(\mathcal{A}) = \{(e, e) | e \in \mathcal{A}\}$
- The relative product or composition of two binary relations \mathcal{R} and \mathcal{R}' is $\mathcal{R} \circ \mathcal{R}' = \{(e, e') | \exists t \in \mathcal{Y} : ((e, t) \in \mathcal{R}) \& ((t, e') \in \mathcal{R}')\}$.
- The inverse of the relation \mathcal{R} is $\mathcal{R}^{-1} = \{(e, e') | (e', e) \in \mathcal{R}\}$.
- The set of images of e is defined by $e.\mathcal{R} = \{e' | (e, e') \in \mathcal{R}\}$.
- The cardinality of \mathcal{R} is defined by $Card(\mathcal{R}) =$ the numbers of pairs in \mathcal{R} .
- The domain of \mathcal{R} is defined by $Dom(\mathcal{R}) = \{e | \exists e' : (e, e') \in \mathcal{R}\}$.
- The range or codomain of \mathcal{R} is defined by $Cod(\mathcal{R}) = \{e' | \exists e : (e, e') \in \mathcal{R}\}$.

Formal Concept Analysis Formal Concept Analysis (FCA) [5, 6] is a mathematical theory of data analysis using formal contexts and concept lattices. It was introduced by Rudolf Wille in 1984, and builds on applied lattice and order theory that were developed by Birkhoff et al. [7]

Definition 1. Formal context: A formal context (or an extraction context) is a triplet $\mathcal{K} = (\mathcal{X}, \mathcal{Y}, \mathcal{R})$, where \mathcal{X} represents a finite set of objects, \mathcal{Y} is a finite set of attributes (or properties) and \mathcal{R} is a binary (incidence) relation (i.e., $\mathcal{R} \subseteq \mathcal{X} \times \mathcal{Y}$). Each couple $(x, y) \in \mathcal{R}$ expresses that the object $x \in \mathcal{X}$ contains the item $y \in \mathcal{Y}$.

Consider the following cross-table (input data, taken from [8]). Let $\mathcal{X} = \{\text{leech}, \text{bream}, \text{frog}, \text{dog}, \text{spike} - \text{weed}, \text{reed}, \text{bean}, \text{maize}\}$ be a set of objects and $\mathcal{Y} = \{\mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{e}, \mathbf{f}, \mathbf{g}, \mathbf{h}, \mathbf{i}\}$ be a set of the properties as defined in Table 1.

b: lives in water,	c: lives on land,
d: needs chlorophyll to produce food,	e: two seed leaves,
f: one seed leaf,	g: can move around,
h: has limbs,	i: suckles its offspring.

Table 1. Properties Description

Let \mathcal{R} reflects the binary relation depicted in Table 2 which can be interpreted as an association between sentences (elements of \mathcal{X}) and words (elements of \mathcal{Y}).

	b	g	h	c	i	d	f	e
$s_0 = \text{leech}$	1	1	0	0	0	0	0	0
$s_1 = \text{bream}$	1	1	1	0	0	0	0	0
$s_2 = \text{frog}$	1	1	1	1	0	0	0	0
$s_3 = \text{dog}$	0	1	1	1	1	0	0	0
$s_4 = \text{spike} - \text{weed}$	1	0	0	0	0	0	1	1
$s_5 = \text{reed}$	1	0	0	1	0	1	1	0
$s_6 = \text{bean}$	0	0	0	1	0	1	0	1
$s_7 = \text{maize}$	0	0	0	1	0	1	1	0

Table 2. Binary Relation \mathcal{R}

Example 1. If we consider that the attribute **b** has a maximum weight, we split the working binary relation in two sub-relations: The first contains all rows validating the attribute **b** and a second sub-relations where the attribute **b** is not valid. Hence, we present the following example:

Definition 2. (HYPER RECTANGLE)

Let $(\mathcal{X}, \mathcal{Y}, \mathcal{R})$ be a formal context and $a \in \mathcal{Y}$, an arbitrary attribute. The Hyper Rectangle, denoted by \mathcal{H}_a , is a sub-relation of \mathcal{R} such that $\mathcal{H}_a(\mathcal{R}) = \mathcal{I}(a.\mathcal{R}^{-1}) \circ \mathcal{R}$. Since the HYPER RECTANGLE, as a binary relation denoted \mathcal{BR} , is related to an attribute a , we define its domain and codomain as follows:

	b	g	h	c	i	d	f	e		b	g	h	c	i	d	f	e
s_0	1	1	0	0	0	0	0	0									
s_1	1	1	1	0	0	0	0	0	s_3	0	1	0	1	0	0	0	1
s_2	1	1	1	1	0	0	0	0	s_6	0	0	1	1	0	1	1	0
s_4	1	1	0	0	0	0	0	0	s_7	0	0	1	1	0	1	0	0
s_5	1	0	0	0	0	1	1	0									

Table 3. (Right): The Hyper Rectangle associated to the attribute b **(Left):** The Remaining Binary relation associated to the attribute b

- $Dom(\mathcal{H}_a(\mathcal{R})) = \mathcal{H}_a(\mathcal{R}).a = \{e \in \mathcal{X} | (e, a) \in \mathcal{H}_a(\mathcal{R})\}$.
- $Cod(\mathcal{H}_a(\mathcal{R})) = \{y \in \mathcal{Y} | (e, y) \in \mathcal{H}_a(\mathcal{R})\}$,
 $= \{e \in \mathcal{X} | (e, a) \in \mathcal{H}_a(\mathcal{R})\}$.
 $= \bigcup \{y \in \mathcal{Y} | (e, y) \in \mathcal{H}_a(\mathcal{R}) \text{ and } e \in Dom(\mathcal{H}_a(\mathcal{R}))\}$.

For each HYPER RECTANGLE, we associate a weight which measures its strengthen in terms of associations between objects and properties. The maximal weight attribute is the more general one since it is shared (directly or indirectly) by the majority of objects. A formalization of this weight is given as below.

Definition 3. (HYPER RECTANGLE WEIGHT)

Let $\mathcal{H}_a(\mathcal{R})$ a Hyper Rectangle associated to an attribute a. The weight $\mathcal{W}(\mathcal{H}_a(\mathcal{R}))$ of the Hyper Rectangle $\mathcal{H}_a(\mathcal{R})$ is defined by a significant information optimization criteria used in the former work of [9]. The economy of a binary relation or the gain in storage space can be calculated in accordance to the following equation:

$$\mathcal{W}(\mathcal{H}_a(\mathcal{R})) = (r/(d * c)) * (r - (d + c)) \quad (1)$$

where r is the cardinality of $\mathcal{H}_a(\mathcal{R})$ (i.e. the number of pairs in binary relation $\mathcal{H}_a(\mathcal{R})$), d is the cardinality of $Dom(\mathcal{H}_a(\mathcal{R}))$, and c is the cardinality of $Cod(\mathcal{H}_a(\mathcal{R}))$.

Example 2. Let's consider the binary relation from Table 2. For each attribute $x \in \mathcal{Y}$, we can extract a Hyper Rectangle $\mathcal{H}_x(\mathcal{R})$ and compute its corresponding weight as presented in Table 4.

\mathcal{W}_b	\mathcal{W}_g	\mathcal{W}_h	\mathcal{W}_c	\mathcal{W}_i	\mathcal{W}_d	\mathcal{W}_f	\mathcal{W}_e
2.667	2.6	2.2	2.25	-1.0	2.6	2.5	-1.0

Table 4. Attributes Weight of the Hyper Rectangle

Definition 4. (OPTIMAL HYPER RECTANGLE)

Let $\mathcal{H}_a(\mathcal{R})$ be a Hyper Rectangle associated to an attribute a. $\mathcal{H}_a(\mathcal{R})$ is said optimal Hyper Rectangle, denoted by $max\mathcal{H}_a(\mathcal{R})$, if and only if $\mathcal{W}_a(\mathcal{H}_a(\mathcal{R})) > \mathcal{W}_y(\mathcal{H}_y(\mathcal{R})) \forall y \neq a, y \in Cod(\mathcal{R})$.

Definition 5. (REMAINING BINARY RELATION)

Let $\mathcal{H}_a(\mathcal{R})$ be a *Hyper Rectangle* associated to the attribute a . The remaining binary relation is the relation \mathcal{R} minus the maximal HYPER RECTANGLE. Thus, we define this remaining binary relation as: $\mathcal{R}_m(\mathcal{R}) = \mathcal{R} - \max\mathcal{H}_a(\mathcal{R})$.

The relation $\mathcal{R}_m(\mathcal{R})$ plays an important role in the construction of the HYPER RECTANGLE coverage. In fact, from the remaining relation, we extract the HYPER RECTANGLES according to the attribute weight until obtaining $\mathcal{R}_m(\mathcal{R})$ as a HYPER RECTANGLE.

Remark 1. For the OPTIMAL HYPER RECTANGLE and all HYPER RECTANGLES that may be extracted from the $\mathcal{R}_m(\mathcal{R})$, we obtain a coverage of \mathcal{R} . From this coverage, we select the word behind each HYPER RECTANGLE to be in the next level of the browsing tree. As \mathcal{BR} , each HYPER RECTANGLE can be explored itself by the same extraction process (after removing its associate attribute). Hence, we build a more specific level of attributes associated to it. Recursively, we can build a browsing tree reflecting different information levels useful for data structuring/browsing.

Example 3. In Figure 1, we present the Hyper rectangles tree extracted from the binary relation depicted in Table 2.

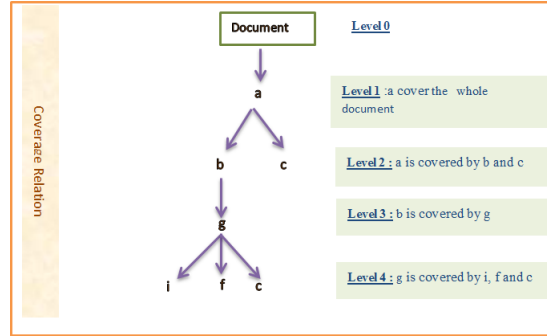


Fig. 1. n-ary tree generation

3 HYPER RECTANGLES COVERAGE

The generation of the attributes, based on the *Hyper Rectangle* and its remaining sub-relations, ensure a coverage of the textual documents at a given level i . So, for a given level i , we compute the attributes for the level $i+1$ with a conservation of the coverage of the previous level. These attributes are browsed as a structure of a tree.

4 Conclusion and Future Work

This system employs some hyper Rectangle based on formal concept analysis as an approach for knowledge discovery and clustering. A heuristic process of finding coverage of the domain of knowledge using the idea of concepts [8] is here replaced by hyper Rectangle ordered in decreasing importance of generated formal words. A good discrimination of the expanding approach consists of generating the pertinent words in decreasing order of importance which have not been considered in the previous approach. Our approach is experimented for text structuring, and it will be used to classify a list of documents in a given corpus. The presented methods may also be used as a base to improve proposed heuristics for solving the NP-complete problem of binary relation coverage with a minimal number of formal words. As a perspective, we envisage to explore the incremental version of this approach and by the inclusion of the similarity between words in the textual document.

Acknowledgment

This research work is supported by Financial Watch (QNRF) project. This publication was made possible by a grant from the Qatar National Research Fund NPRP085831101. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the QNRF.

References

1. Godin, R., Missaoui, R., Alaoui, H.: Learning algorithms using a galois lattice structure. In: Proceedings of third IEEE International Conference on Tools for AI, San Jose, CA, USA. (1991) 22–29
2. Carpineto, C., Romano, G.: Exploiting the potential of concept lattices for information retrieval with credo. *Journal of Universal Computing* (2004) 10, 8,985–1013
3. Jaoua, A., Al-Saidi, M., Hasnah, A., Jaam, J., Ahmed, S., Salem, B., Rashid, N., Shareef, S., Zaghlan, S.: Structured conceptual meta-search engine, fourth international conference on concept lattices an applications. CLA2006 (2006)
4. Schmidt, G., Stroehlein, T.: *Relations and Graphs: Discrete Mathematics for Computer Scientists*. Springer-Verlag (1993)
5. Ganter, B.: Two basic algorithms in concept analysis. Preprint 831, Technische Hochschule, Darmstadt, Germany (1984)
6. Wang, L.: Fuzzy systems and knowledge discovery. In: Proceedings Part I of the Second International Conference, FSKD 2005), Changsha, China. Volume ISBN 10 3-540-28312-9., Springer-Verlag berlin Heidelberg 2005 (2005) 515–519
7. G.Birkhoff: *Lattice Theory*, 1st edn. Providence: Amer. Math. Soc. (1965)
8. Ganter, B., Wille, R.: *Formal Concept Analysis: mathematic Foundations*. Springer-Verlag (1999)
9. Jaoua, A.: Pseudo_conceptual text and web strunturing in the third conceptual structures tool interoperability workshop. 16th International Conference on Conceptual Structures(ICCS 2008), Toulouse, France (2008)