جامعة قطر
QATAR UNIVERSITY

Hyper-Rectangular Relation Decomposition

Application for Text Mining and Structuring

Prepared by: Fethi FERJANI

Agenda

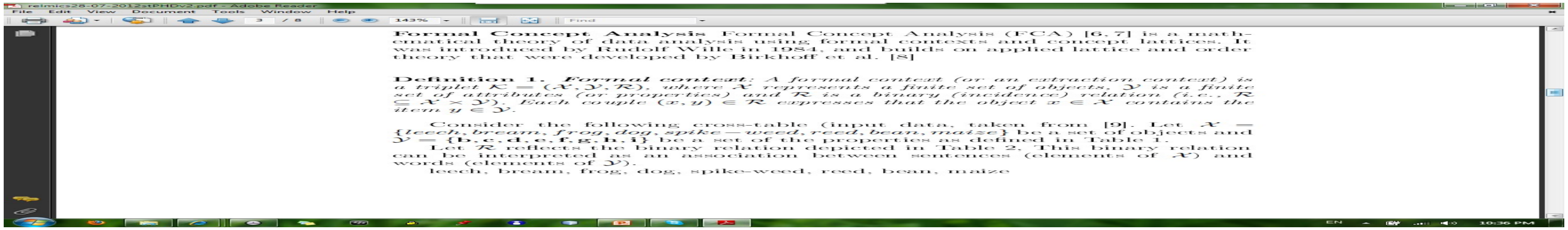
1. **Problem and approach**
2. **About Hyper Rectangles**
3. **Hyper Rectangular Decomposition and Coverage**
4. **Conclusion and future work**

Problem and Approach

- **What**
  - Achieve structuring and browsing capabilities in a large-scale corpus on two different levels:
    - Macro-level browsing: deals with the global visual and logical structure of the corpus (e.g. categories, sub-categories, sections, fields, etc).
    - Micro-level browsing: is used for navigating through a textual document. For example, company information, location, etc) .
- **How**
  - New method handles construction of a Hyper Rectangular Generic tree (GT).
  - Introducing scalability and strong sematic structures.
  - Introduces a solution to the rectangles labeling challenge

Hyper Rectangle Definition

Formal Context


reimics28-07-2012stPHDv2.pdf - Adobe Reader
File Edit View Document Tools Window Help
3 / 8
143%
Find
EN
10:36 PM

Formal Concept Analysis Formal Concept Analysis (FCA) [6, 7] 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. [8]

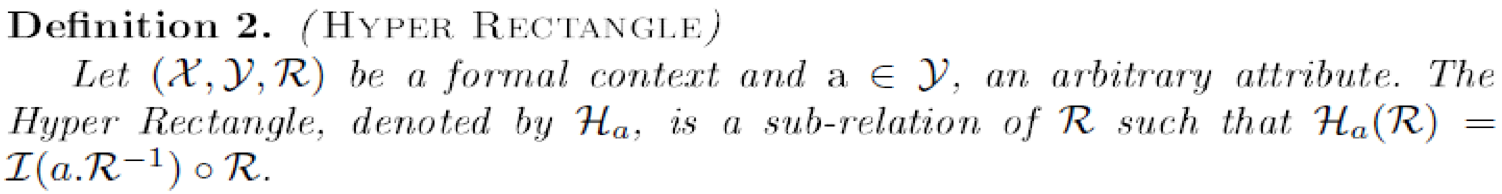
Definition 1. ***Formal context:*** *A formal context (or an extraction context) is a triplet $\mathcal{K} = (\mathcal{X}, \mathcal{Y}, \mathcal{R})$, where $\mathcal{X}$ represents a finite set of objects, $\mathcal{Y}$ is a finite set of attributes (or properties) and $\mathcal{R}$ is a binary (incidence) 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 [9]. Let $\mathcal{X} = \{leech, bream, frog, dog, spike-weed, reed, bean, maize\}$ 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.

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

leech, bream, frog, dog, spike-weed, reed, bean, maize

Hyper Rectangle



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 $\mathcal{H}_a(\mathcal{R}) = \mathcal{I}(a.\mathcal{R}^{-1}) \circ \mathcal{R}$.*

- Hyper Rectangle is the union of all possible rectangles in a binary relation involving an element of the domain or the range of a given binary relation

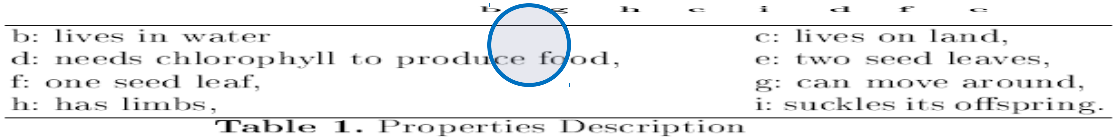
Example of Hyper Rectangle

b: lives in water
d: needs chlorophyll to produce food,
f: one seed leaf,
h: has limbs,

c: lives on land,
e: two seed leaves,
g: can move around,
i: suckles its offspring.

Table 1. Properties Description

b g h c i d f e

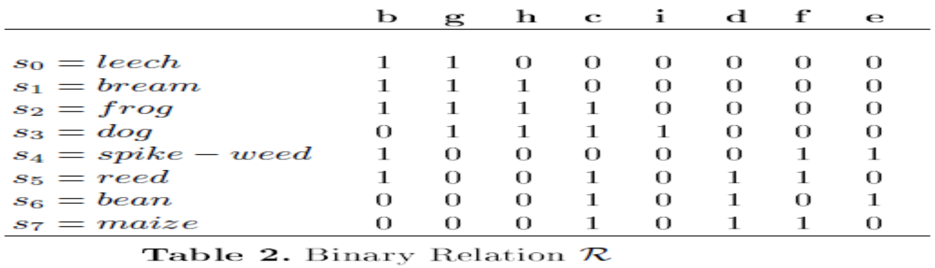


b: lives in water
d: needs chlorophyll to produce food,
f: one seed leaf,
h: has limbs,

c: lives on land,
e: two seed leaves,
g: can move around,
i: suckles its offspring.

Table 1. Properties Description

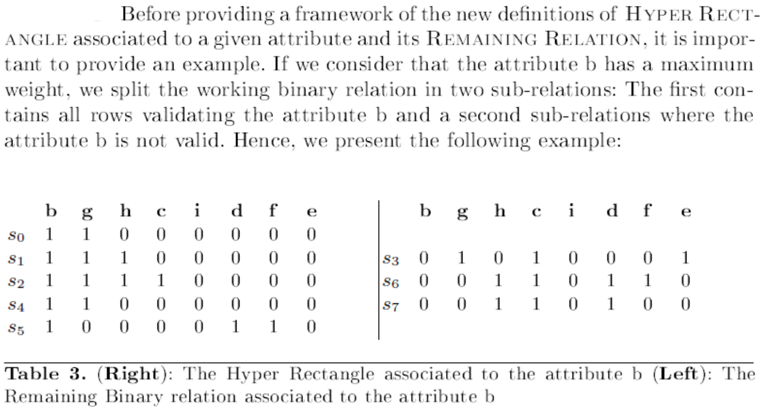
Binary Relation R



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

Table 2. Binary Relation $\mathcal{R}$

Hyper Rectangle Hb(R)



Before providing a framework of the new definitions of Hyper Rectangle associated to a given attribute and its Remaining Relation, it is important to provide an example. 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:

	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_2	1	1	1	1	0	0	0	0
s_4	1	1	0	0	0	0	0	0
s_5	1	0	0	0	0	1	1	0

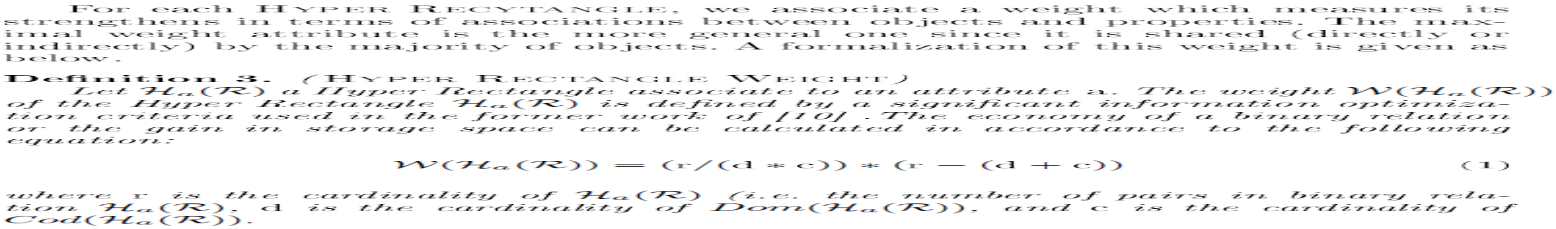
	b	g	h	c	i	d	f	e
s_3	0	1	0	1	0	0	0	1
s_6	0	0	1	1	0	1	1	0
s_7	0	0	1	1	0	1	0	0

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

Hyper Rectangle Weighting

- Optimization criterion for the decomposition process
- Associates a strength measure to each property
- Higher weight indicates generality of an attribute as it is shared with more objects

- **Hyper Rectangle Weight**

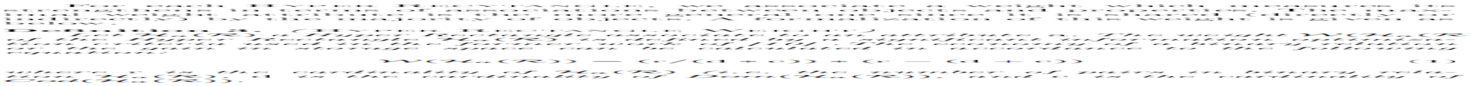


For each Hyper Recytangle, we associate a weight which measures its strengthens in terms of associations between objects and properties. The maximal weight attribute is the more general one since it is shared (directly or indirectly) by the majority of objects. A formalization of this weight is given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

Definition 3. (Hyper Rectangle Weight)
Let $\mathcal{H}_a(\mathcal{R})$ a Hyper Rectangle associate 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 [10] . 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 $Cod(\mathcal{H}_a(\mathcal{R}))$.*



For each Hyper Recytangle, we associate a weight which measures its strengthens 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.

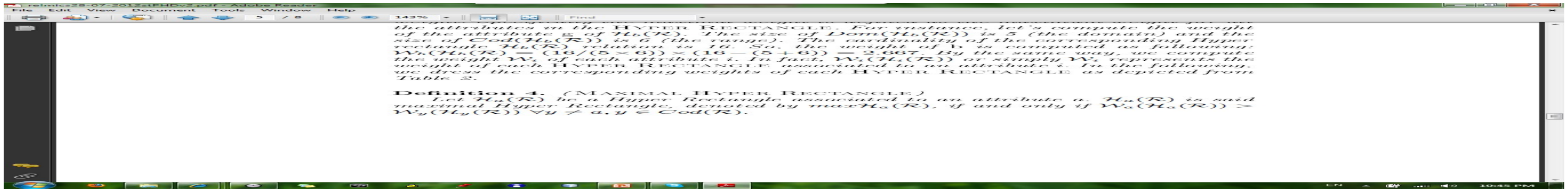
Definition 3. (Hyper Rectangle Weight) *Let $\mathcal{H}_a(\mathcal{R})$ a Hyper Rectangle associate to an attribute a. The weight $\mathcal{W}(\mathcal{H}_a(\mathcal{R}))$ of the Hyper Rectangle $\mathcal{H}_a(\mathcal{R})$ is defined by a significant information optimization criteria used in the former work of [10] . 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 $Cod(\mathcal{H}_a(\mathcal{R}))$.*

Maximal Hyper Rectangles

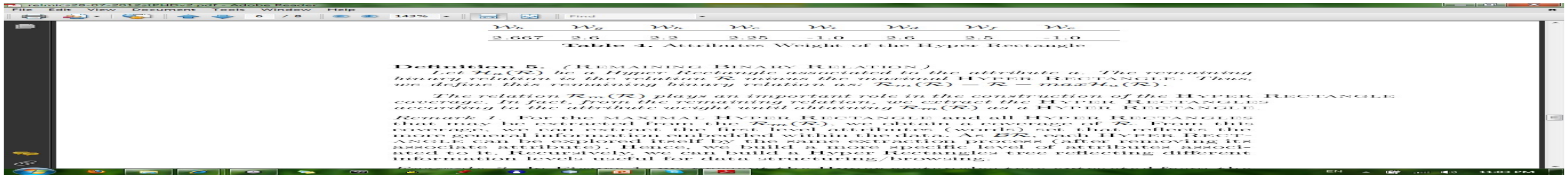
- **Maximal Hyper Rectangle**


relmics28-07-2012stPHDv2.pdf - Adobe Reader
File Edit View Document Tools Window Help
5 / 8
143%
Find

the attributes in the Hyper Rectangle. For instance, let's compute the weight of the attribute g of $\mathcal{H}_b(\mathcal{R})$. The size of $Dom(\mathcal{H}_b(\mathcal{R}))$ is 5 (the domain) and the size of $Cod(\mathcal{H}_b(\mathcal{R}))$ is 6 (the range). The cardinality of the corresponding Hyper rectangle $\mathcal{H}_b(\mathcal{R})$ relation is 16. So, the weight of b is computed as following: $\mathcal{W}_b(\mathcal{H}_b(\mathcal{R})) = (16/(5\times6))\times(16-(5+6)) = 2.667$. By the same way, we compute the weight $\mathcal{W}_i$ of each attribute i. In fact, $\mathcal{W}_i(\mathcal{H}_i(\mathcal{R}))$ or simply $\mathcal{W}_i$ represents the weight of each Hyper Rectangle associated to an attribute i. In the following, we dress the corresponding weights of each Hyper Rectangle as depicted from Table 2.

Definition 4. (Maximal Hyper Rectangle) *Let $\mathcal{H}_a(\mathcal{R})$ be a Hyper Rectangle associated to an attribute a. $\mathcal{H}_a(\mathcal{R})$ is said maximal Hyper Rectangle, denoted by $max\mathcal{H}_a(\mathcal{R})$, if and only if $\mathcal{W}_a(\mathcal{H}_a(\mathcal{R})) > \mathcal{W}_y(\mathcal{H}_y(\mathcal{R})) \forall y \neq a, y \in Cod(\mathcal{R})$.*

- **Remaining Binary Relation**



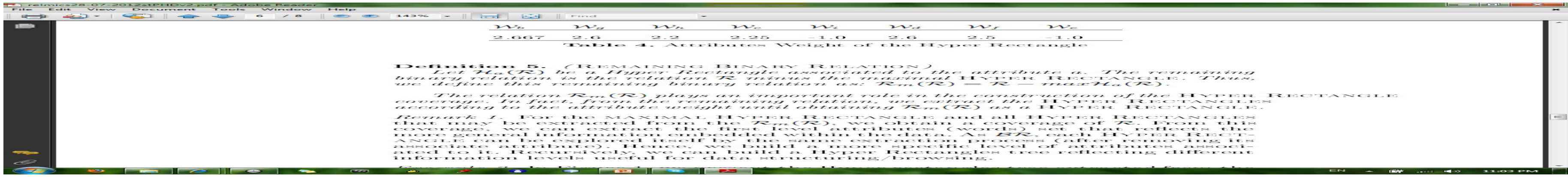
$\mathcal{W}_b$	$\mathcal{W}_g$	$\mathcal{W}_h$	$\mathcal{W}_c$	$\mathcal{W}_i$	$\mathcal{W}_d$	$\mathcal{W}_f$	$\mathcal{W}_e$
2.667	2.6	2.2	2.25	-1.0	2.6	2.5	-1.0

Table 4. Attributes Weight of the Hyper Rectangle

Definition 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 maximal 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 can extract the first level attributes (words) set that reflects the more general information embedded within the data. As *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 Hyper Rectangles tree reflecting different information levels useful for data structuring/browsing.


relmics28-07-2012stPHDv2.pdf - Adobe Reader
File Edit View Document Tools Window Help
6 / 8
143%
Find

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

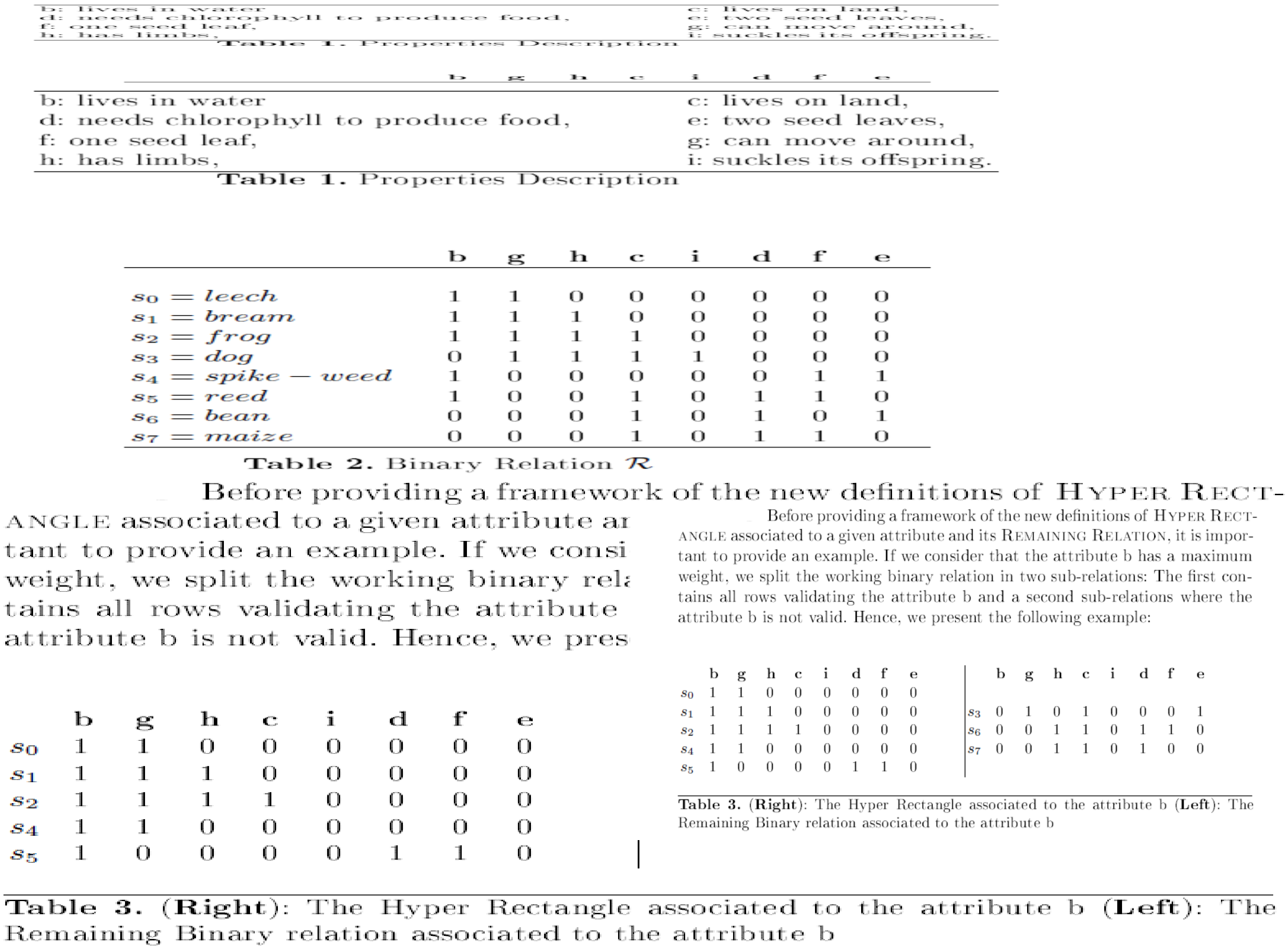
Table 4. Attributes Weight of the Hyper Rectangle

Definition 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 maximal 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 can extract the first level attributes (words) set that reflects the more general information embedded within the data. As *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 Hyper Rectangles tree reflecting different information levels useful for data structuring/browsing.

Maximal Hyper Rectangle and Remaining Relation



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

Table 1. Properties Description

b g h c i d f e

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

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

Table 2. Binary Relation $\mathcal{R}$

Before providing a framework of the new definitions of HYPER RECT-
ANGLE associated to a given attribute ar
tant to provide an example. If we consi
weight, we split the working binary rel
tains all rows validating the attribute
attribute b is not valid. Hence, we pres

Before providing a framework of the new definitions of HYPER RECTANGLE associated to a given attribute and its REMAINING RELATION, it is important to provide an example. 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:

	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_2	1	1	1	1	0	0	0	0
s_4	1	1	0	0	0	0	0	0
s_5	1	0	0	0	0	1	1	0

	b	g	h	c	i	d	f	e
s_3	0	1	0	1	0	0	0	1
s_6	0	0	1	1	0	1	1	0
s_7	0	0	1	1	0	1	0	0

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

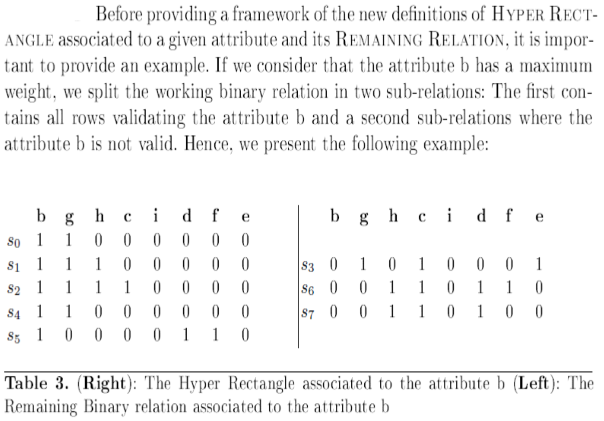
	b	g	h	c	i	d	f	e
s_0	1	1	0	0	0	0	0	0
s_1	1	1	1	0	0	0	0	0
s_2	1	1	1	1	0	0	0	0
s_4	1	1	0	0	0	0	0	0
s_5	1	0	0	0	0	1	1	0

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

Hyper Rectangular Coverage

- Calculate the weight for all attributes and sort them descending
- Perform generalization of the best attribute into Hyper Rectangle
- Find the Hyper rectangles in the remaining relation recursively until the entire relation is covered or Rm(R) is a Hyper Rectangle
- For a given level **i**, we compute the attributes for the level **i+1** with a conservation of the coverage of the previous level.

Generic Tree of words

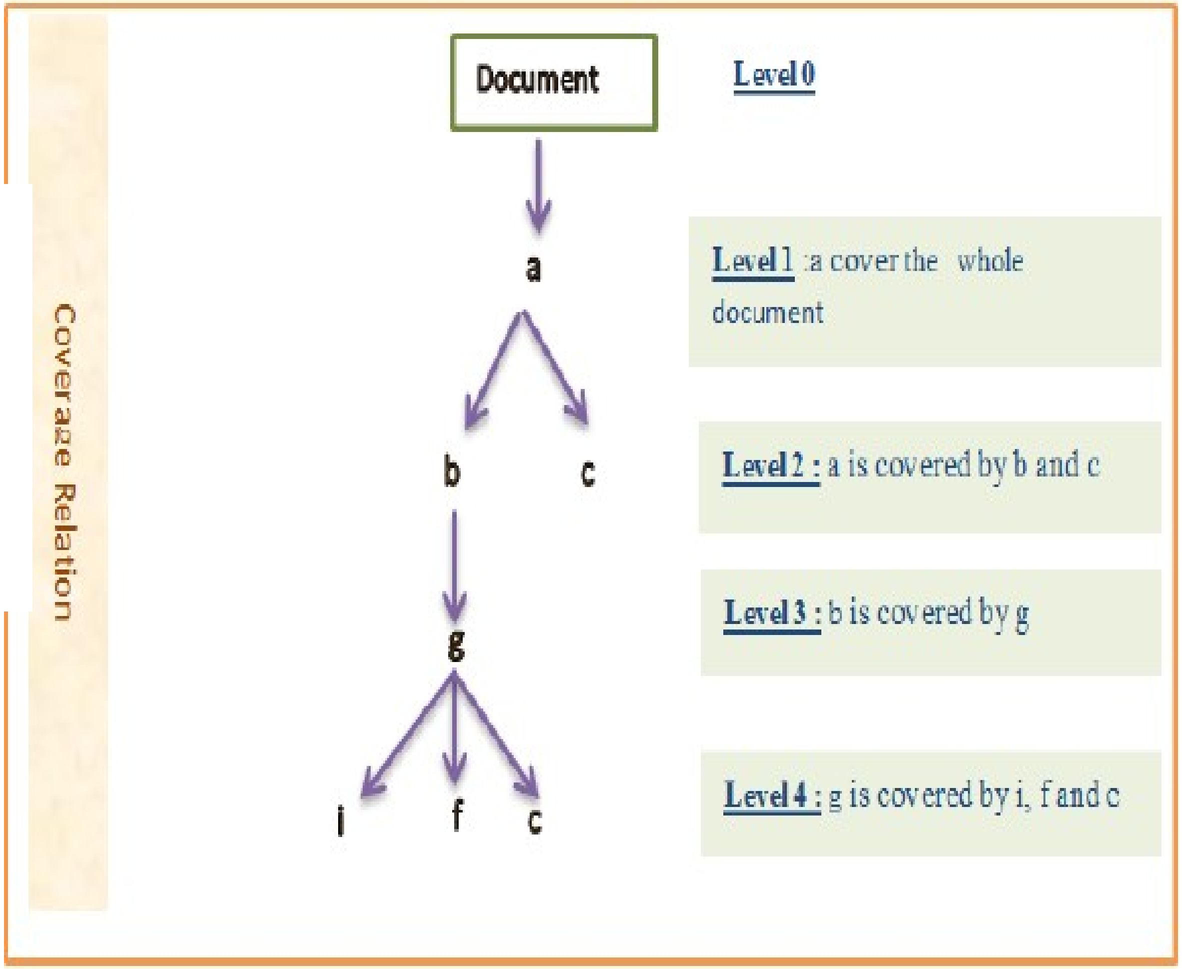


Before providing a framework of the new definitions of Hyper Rectangle associated to a given attribute and its Remaining Relation, it is important to provide an example. If we consider that the attribute b has a maximum weight, we split the working binary relation in two sub-relations: The first contains all rows validating the attribute b and a second sub-relations where the attribute b is not valid. Hence, we present the following example:

	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_2	1	1	1	1	0	0	0	0
s_4	1	1	0	0	0	0	0	0
s_5	1	0	0	0	0	1	1	0

	b	g	h	c	i	d	f	e
s_3	0	1	0	1	0	0	0	1
s_6	0	0	1	1	0	1	1	0
s_7	0	0	1	1	0	1	0	0

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



Level 0

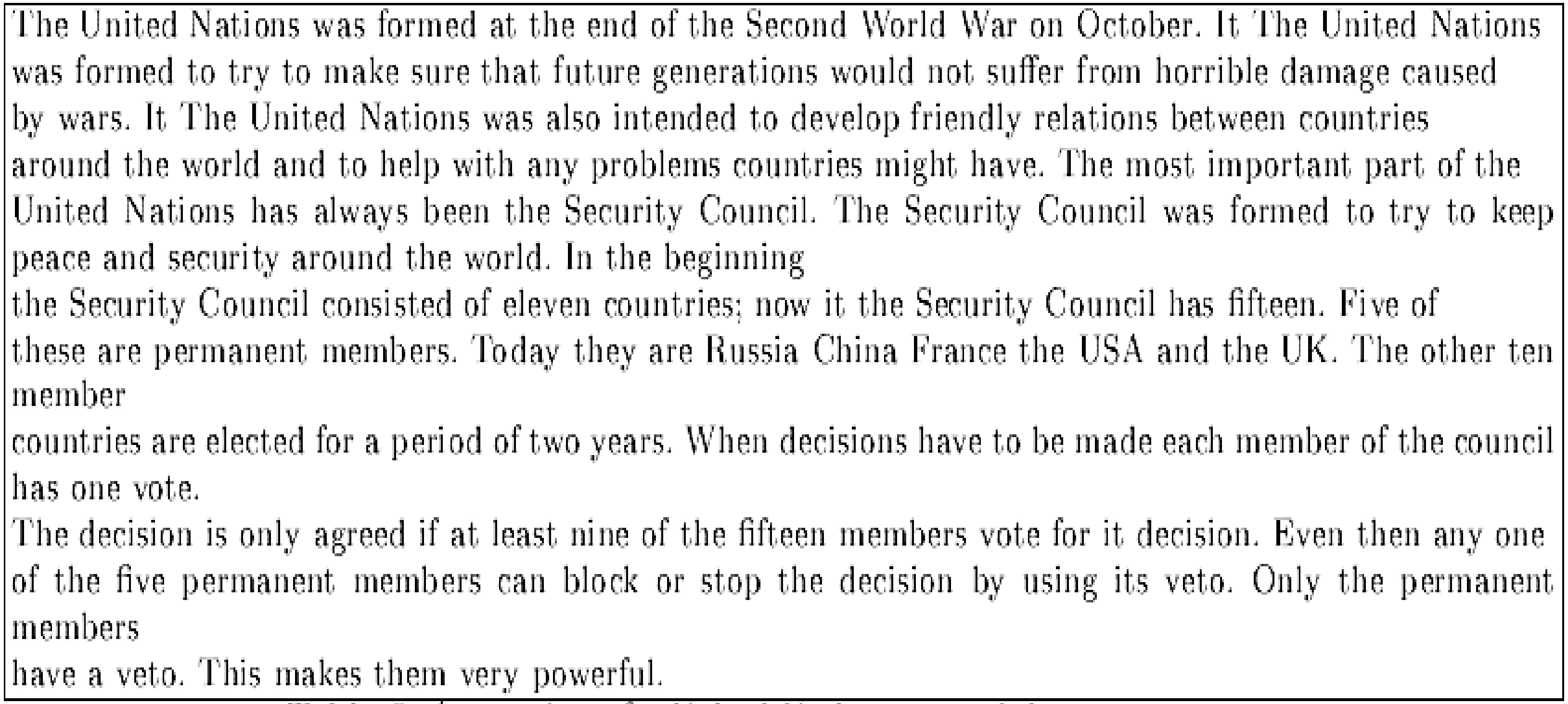
Level 1 :a cover the whole document

Level 2 : a is covered by b and c

Level 3 : b is covered by g

Level 4 : g is covered by i, f and c

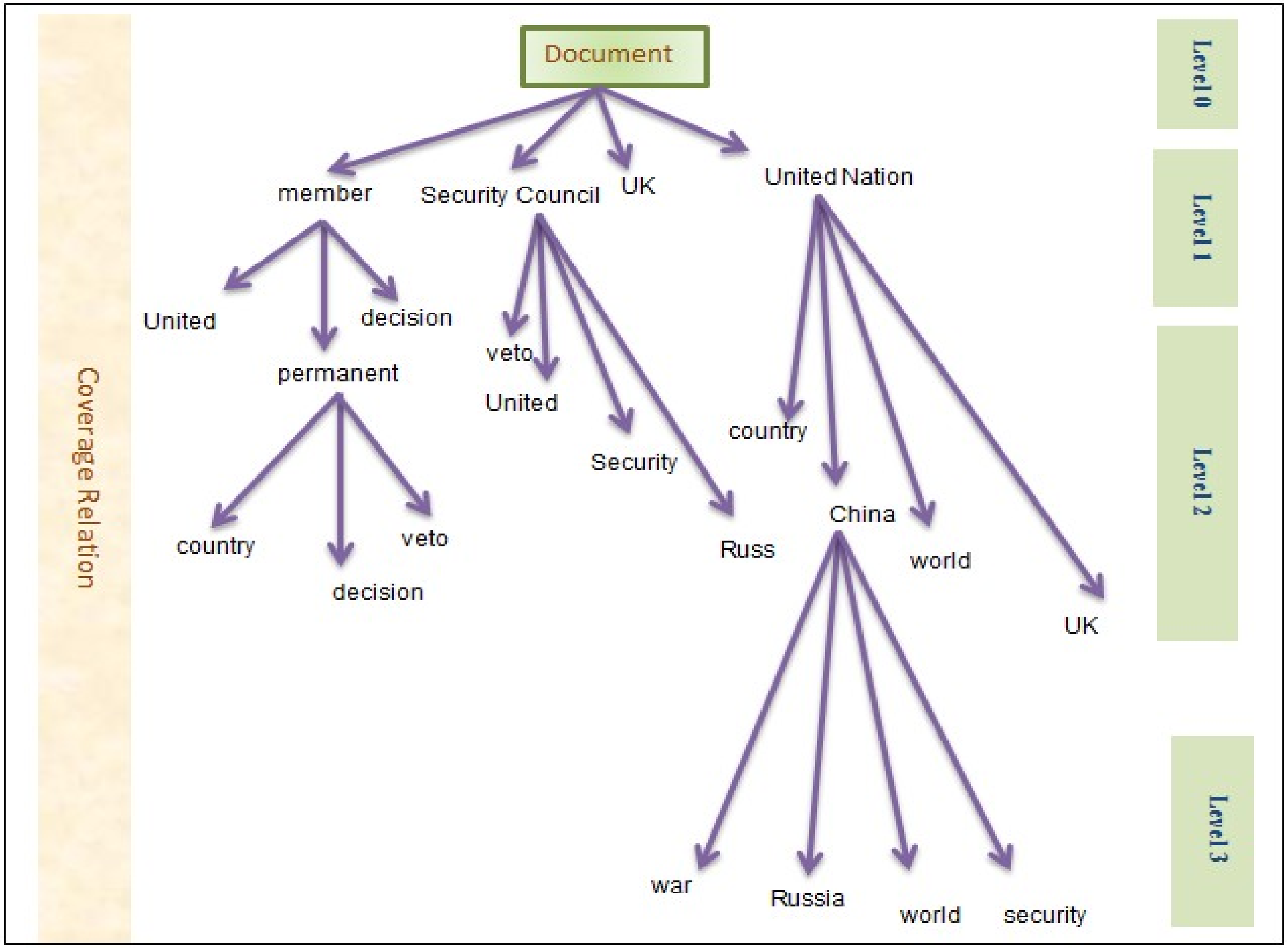
Illustrative Example United Text document



The United Nations was formed at the end of the Second World War on October. It The United Nations was formed to try to make sure that future generations would not suffer from horrible damage caused by wars. It The United Nations was also intended to develop friendly relations between countries around the world and to help with any problems countries might have. The most important part of the United Nations has always been the Security Council. The Security Council was formed to try to keep peace and security around the world. In the beginning the Security Council consisted of eleven countries; now it the Security Council has fifteen. Five of these are permanent members. Today they are Russia China France the USA and the UK. The other ten member countries are elected for a period of two years. When decisions have to be made each member of the council has one vote. The decision is only agreed if at least nine of the fifteen members vote for it decision. Even then any one of the five permanent members can block or stop the decision by using its veto. Only the permanent members have a veto. This makes them very powerful.

Table 6. A snapshot of a United Nations textual document.

Generic Tree Correspondent



Conclusion

- New heuristic based approach utilizing Hyper Rectangles for navigating and browsing inside a corpus was presented

- Hyper Rectangle is the union of all possible rectangles in a binary relation involving an element of the domain or the range of a given binary relation

- Hyper Rectangles are very appropriate for large data sets and may overcome the previous approaches drawbacks regarding scalability

- 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.

جامعة قطر
QATAR UNIVERSITY

Thank You