CBKR+: A Conceptual Framework for Improving Corpus Based Knowledge Representation

by

Shabnam Surjitsingh Ivković

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2006

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In Corpus Based Knowledge Representation [CBKR], limited association capability, that is, no criteria in place to extract substantial associations in the corpus, and lack of support for hypothesis testing and prediction in context, restricted the application of the methodology by information specialists and data analysts. In this thesis, the researcher proposed a framework called CBKR+ to increase the expressiveness of CBKR by identifying and incorporating association criteria to allow the support of new forms of analyses related to hypothesis testing and prediction in context.

As contributions of the CBKR+ framework, the researcher (1) defined a new domain categorization model called Basis for Categorization model, (2) incorporated the Basis for Categorization model to (a) facilitate a first level categorization of the schema components in the corpus, and (b) define the Set of Criteria for Association to cover all types of associations and association agents, (3) defined analysis mechanisms to identify and extract further associations in the corpus in the form of the Set of Criteria for Association, and (4) improved the expressiveness of the representation, and made it suitable for hypothesis testing and prediction in context using the above.

The application of the framework was demonstrated, first, by using it on examples from the CBKR methodology, and second, by applying it on 12 domain representations acquired from multiple sources from the physical-world domain of Criminology. The researcher arrived at the conclusion that the proposed CBKR+ framework provided an organized approach that was more expressive, and supported deeper analyses through more diagnostic and probability-based forms of queries.

Acknowledgements

I thank my advisor, Professor Anne Banks Pidduck, who patiently and efficiently oversaw the completion of my thesis. In addition to Occam's Razor, her words 'Keep things simple' helped me to focus my research and give it value. Her help and guidance were of key importance to the completion of my Master's degree.

I extend my sincere gratitude to Professor Paulo Alencar, for his time, his understanding of this work as a reader, and the invaluable support and feedback I've received from him during the entire course of my degree. I express my deep appreciation to Professor Don Cowan, for his insight and suggestions as a reader of this thesis. I thank Professor Frank Tompa for his resources that made this degree possible, and Professor Robin Cohen for her insight and support of this research. I extend my deepest gratitude to Suzanne Safayeni and Margaret Towell for their tireless encouragement and support. Lastly, I'd like to thank Candace Newman, Brenda McBay, Wendy Rush and Jessica Miranda for being wonderful and encouraging friends.

My deepest gratitude goes to my wonderful husband, Igor. He is my strength when I am weak and my support whenever I need it. I thank him for believing in me and not letting me ever give up. I express my earnest gratefulness to my parents, Papa; who always insisted that anything worth doing is worth doing well or not at all, and Mama; who still keeps reminding me that if I do not stand up for something, I will fall for anything. I extend my heartfelt appreciation to my wonderful family, Mummy, Daddy, Sandra and Baka for their unequivocal encouragement and untiring support. They have stood steadfastly by me in the most challenging of times, and showered me with unconditional love.

Knowledge and wisdom, far from being one,
Have oft times no connexion.
Knowledge dwells
In heads replete with thoughts of other men;
Wisdom in minds attentive to their own.
Knowledge, a rude unprofitable mass,
The mere material with which Wisdom builds,
Till smooth'd and squar'd, and fitted to its place,
Does but encumber what it means to enrich.
Knowledge is proud that he has learn'd so much,
Wisdom is humble that he knows no more.

William Cowper (1731-1800), The Task

Table of Contents

List of Figures and Tables

Chapter 1 – Introduction

> *To each individual the world will take on a different connotation of meaning.*
> *The importance lies in the desire to search for an answer.*
> *~ Thomas Stearns Elliot (1888-1965), Poet*

1.0. Chapter Introduction

A knowledge representation consists of a methodology for computer systems to be able to encode and utilize information about the domain under consideration. [Stone, 2003] This methodology must be able to correctly represent the knowledge and the interactions of the domain, and should have the ability to create connections within the components of a domain. One of the more complex aspects of accomplishing the above is that individuals use skills based on cognition and instinct to manipulate information. This manipulation is difficult to simulate in a representation built on a computer based system. Therefore, building models of knowledge representation and reasoning require efficient identification of the elements and the relationships between them, formalization of this information, and computational means to leverage this information. [Stone, 2003]

This chapter introduces some of the problems with the implementation of a knowledge representation system. This is followed by a short discussion on building and evaluating a knowledge representation. The motivation for conducting the research in this thesis is next, followed by the problem definition and the thesis' proposed research contribution.

1.1. Implementation of a Knowledge Representation System

For any domain that is to be represented, three categories of problems have been identified by Woods regarding creation and implementation of the representation. The description of these three problems, as below, serves well to establish the features that a representation must contain [Woods, 1990]:

– *Knowledge acquisition*: Once the domain has been represented, the model should allow the addition of more information about the domain. The model should be able to distinguish between the elements of information, and have a method to deal with those that it cannot resolve. The model should also be able to create generic categories of these elements so that new elements can be easily placed within the model without ambiguity, conflicts and contention. The model must be able to acquire and assimilate knowledge dynamically.

– *Perception*: The generation of a set of possible best-case hypotheses and searching within them should not lead to a combinatorial explosion. At the same time, the model should be able to establish new relationships and interactions between known elements and the new ones, with tolerance of input errors and incorrect hypotheses.
– *Execution*: The model should support monitoring and evaluation of the sharing of large amounts of data of common knowledge between various hypotheses. The model must have efficient contingency plans to re-plan scenarios if things go wrong, or if there are competing expectations of the representation.

Effective knowledge representation of the domain must be able to, in some measure, refine and test generalizations about the data within it and leverage semantic relationships so that the representation can be used for processes such as data mining and hypothesis testing. This gives rise to issues such as 'granularity', or the detail and refinement of the individual elements, and the 'organization' of their relationships with other elements. Since refining the granularity makes the representation more complex, it is advisable that a domain should be optimally refined for the functionality of the application it supports. The organization of relationships deals with formulating patterns, segregating knowledge components with relation to these patterns, and creating categories of general and specific interactions within the domain in a consistent manner with an attempt to cover the domain completely. [Stone, 2003] Supported by some form of categorization, associations and patterns formed within, the representation can be used for indexing of the components of domain information. The efficient application of such a model of representation to larger domains may improve applications such as knowledge mining, question answering, retrieval of data, structured navigation, hypothesis testing, and prediction within the domain. [Aronson & Rindflesch, 1998]

1.2. Building and Evaluating the Representation

Studying the types of knowledge can be substantially helpful while designing the appropriate knowledge representation and association forming schematics for a particular domain with respect to the type of knowledge contained within it.

According to Delugach and Rochowiak [Delugach & Rochowiak, 2001], knowledge can be procedural, declarative, semantic, or episodic. Procedural knowledge deals with the cognitive process of knowing how to perform a task where the knowledge elements are generally sequentially encountered. Another type of knowledge is declarative knowledge, which is essentially explicit knowledge, and is a good place to start the representation of the domain. Semantic knowledge is based on the principles of long-term memory, and it has cognitive structure, where the organized elements comprise words, symbols, meanings, rules, interrelations, and procedures to create semantic relationships. Episodic knowledge comprises knowledge that may be based on biographical and experiential situations, and is acquired through experience that is grouped into episodes.

Building a knowledge representation has been comprehensively described in [Delugach & Rochowiak, 2001]. When building one, there are some steps that have been identified. First, an ontology consisting of things and events of the domain must be generated, as this will provide a stable background for knowledge. A description of a top-level ontological schema may contain the elements, such as objects, processes, relations between objects, localized interaction between objects, purpose of the object within the relations, and combinations of interactions between objects that create situations. Once the ontology has been identified, the representations for operations and for claims can be added based on a predetermined query style. The components of knowledge that help to describe and define the domain consist of propositions about the domain, statements and rules indicating true and false knowledge, procedures for identifying new items, procedures for constructing new items in the ontology, descriptions for various queries that arise on the ontology, and heuristics and rules for solving these queries.

For knowledge to be effective when it is represented, it should be complemented by a mechanism that can generate metaknowledge about the domain by leveraging the knowledge already present in the domain. The mechanism can be based on factors such as inductive or deductive inferences, the mechanism's reliability to continue in the face of ambiguous information, and simplicity and elegance of the mechanism. [Delugach & Rochowiak, 2001]

As an example of how ontological elements, associations and metaknowledge generating mechanisms of the same domain can be related, consider a spaceship being built. The parts or the building blocks have to be put together according to the design plans, using a workforce. Analogically speaking, the parts or the building blocks are the ontological components of the domain, the design plans are the associations that are identified between these components, and the workforce is the metaknowledge generating mechanism that can suggest what supplies are needed and when they should be acquired.

Once built, a knowledge representation must be discussed to judge its effectiveness. A set of criteria, suggested by Torsun, for the study of knowledge representation schemes, is described below [Torsun, 1995 as in Pesonen, 2002]:

- *Semantics:* Any representation should have a semantic theory that helps the representation to correspond to the real physical domain it represents.
- *Expressive adequacy:* Any representation should specifically indicate what knowledge about the domain can be successfully represented, what cannot be, and what knowledge can be conditionally represented.
- *Naturalness:* Language understood by machines is difficult to understand by the user. Though all knowledge is represented in binary form in a computational context, this form will be very difficult for the user to understand and use.
- *Reasoning:* For the complete representation of the domain, it is important to evaluate the extent and the efficiency of the deductive, inductive or abductive processes of the inference mechanism employed by the representation.
- *Primitives:* A knowledge representation must specify and clarify its primitives used in the representation, where primitives are some basic procedures or processes whose working and functionality are not specific to the given representation.
- *Incompleteness:* Due to the complexity of real domains, the knowledge contained within a domain can never be completely represented. Knowledge modeled by a representation is always brittle at the edges of the model. The completeness of the knowledge modeled by the representation system must be evaluated in comparison to the physical world existence of the domain.
- *Revisable reasoning:* Instinctively, and in the human mind, much of what is known about a physical world domain is *almost always* true as the mind can attribute different significances to a single situation. A different significance to a given situation revises it to be *almost always* true in a different given context. The

representation system must be evaluated on its ability to perform such revisable reasoning.

– *Flexibility*: Any given domain is dynamic and susceptible to change. This factor measures the ability to add new pieces of knowledge easily.

1.3. Motivation for Research

Knowledge representation can be complex. A single representation may be unable to represent a domain in all completeness. The initial accumulation and storage of information is relatively easy, while the derivation of knowledge from the accumulated information is a subject of ongoing research. [Sowa, 2000] There are many views on knowledge, leading to various types of representations, each with its own drawbacks and complexity. [Davis et al., 1993] This research is focused on the representation of a macroscopic view of a given domain, via the schemas that represent the structure of the domain. A microscopic view of the domain would be one where every entity and element within the domain, and not just its structure, would be modeled by the representation.

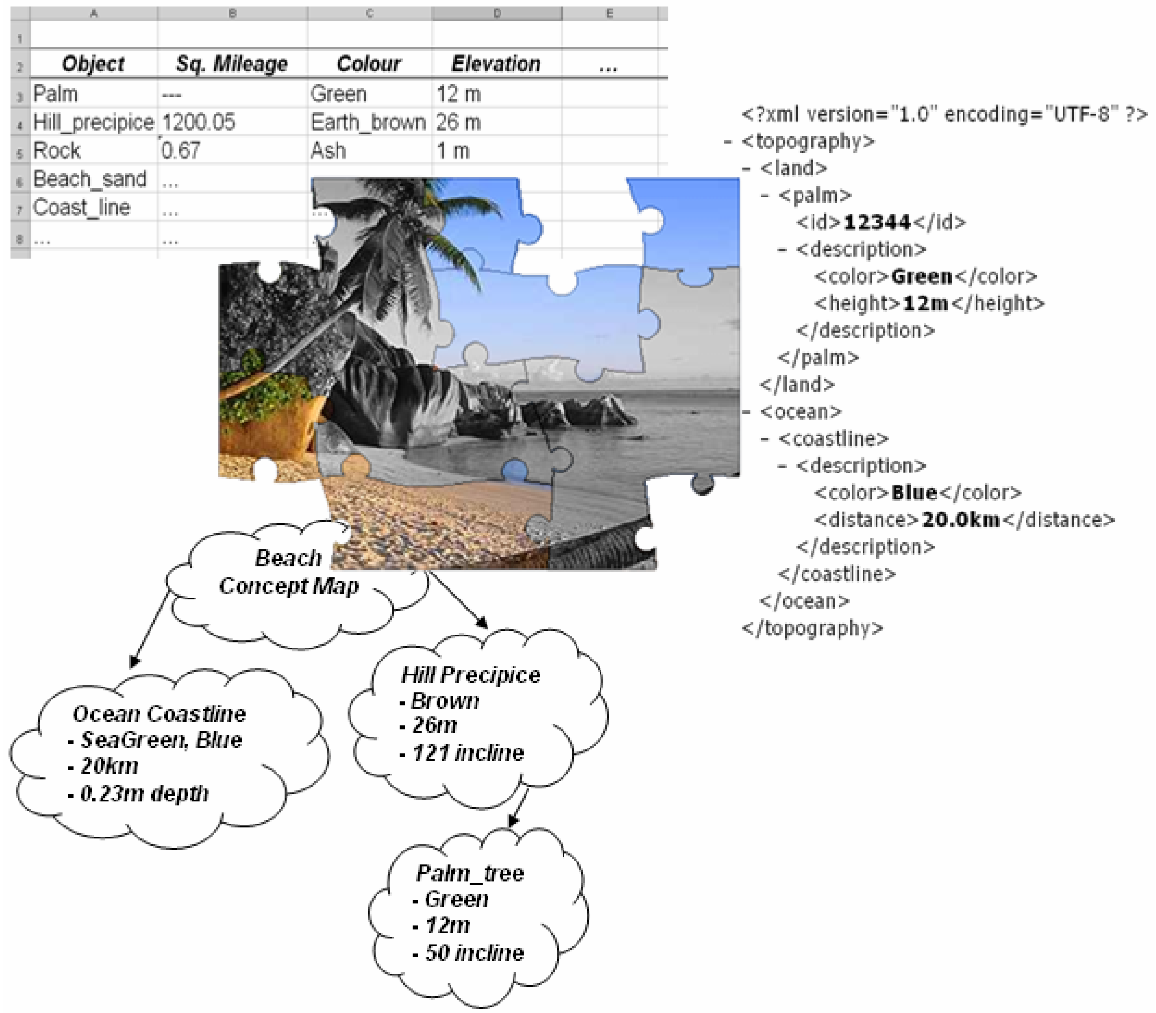
A single domain can be represented by many forms of representation, which can stand for multiple perspectives. The process of building a knowledge representation methodology or framework for a given domain is a complex process. A single schema may not completely model a given domain, given that there may exist numerous perspectives to the domain that cannot all be possibly modeled and leveraged in a single representation. Every domain can be modeled differently depending on the schema used, the representation language employed, or the aspect and perspective of the domain that is best focused on. This makes any representation of a domain brittle in terms of the wholesomeness of the knowledge contained within it.

If a corpus of such schemas is constructed with mappings and associations some measure of the incompleteness of the representation of the domain can be lessened. One methodology that has been proposed to achieve the above is that of Corpus Based Knowledge Representation [CBKR] [Halevy & Madhavan, 2003]. The applications that can be addressed by this model are limited, but the model itself is relatively less complex to implement as compared to traditional knowledge representation methods discussed in the next chapter. Though the CBKR methodology is effective, it is hypothesized and eventually proved over the course of this thesis that this methodology can be improved and made applicable to a wider spectrum of applications. The underlying principle of this improvement is the regularity of occurrence in the associations or mappings between the various representational components of the domain.

1.3.1. What is a Corpus?

Each schema is composed of components that model the domain, and associations between those components. If a library of such schemas is constructed, and mappings, or associations, are established between the various components of the representations, some measure of the incompleteness of the representation of the domain can be alleviated. Such a library of schemas may be called a corpus. [Halevy & Madhavan, 2003] For example, it is difficult to get complete information about the ethos of a country from one book. But a library can have several books, written by different authors, published by different companies, written in different styles, covering different eras, about the same country. All this knowledge put together gives a more comprehensive picture of the ethos of that country.

A corpus indicates a collection of items – it can be a collection of words and grammar rules in Language Processing, a collection of samples in Statistics, a collection of case studies in Psychology, a collection of classes in Software Engineering, and so on. In this work, a corpus is defined as a set of characterizations or characteristics and views of a given domain.



	A	B	C	D	E
1					
2	***Object***	***Sq. Mileage***	***Colour***	***Elevation***	**...**
3	Palm	---	Green	12 m	
4	Hill_precipice	1200.05	Earth_brown	26 m	
5	Rock	0.67	Ash	1 m	
6	Beach_sand	...			
7	Coast_line	...	...		
8	...	...	...		

```
  <?xml version="1.0" encoding="UTF-8" ?>
- <topography>
  - <land>
    - <palm>
        <id>12344</id>
      - <description>
          <color>Green</color>
          <height>12m</height>
        </description>
      </palm>
    </land>
  - <ocean>
    - <coastline>
      - <description>
          <color>Blue</color>
          <distance>20.0km</distance>
        </description>
      </coastline>
    </ocean>
  </topography>
```

Figure 1.1.: Defining a Corpus

In Figure 1.1 the jig-saw puzzle [Jigzone, 2006] stands for the domain that must be represented. One aspect of the domain may be represented by a Concept Map-like schema, another by an XML schema. Yet another aspect may be modeled by a tabular record-like structure. Each of these schemas models the structural level of the domain, where the components of the schemas form a set of characterizations of the domain. This set is called a corpus, and its elements represent the macroscopic structure of a domain. This work considers that corpus-based methodologies fall under the discipline of Knowledge Integration and Management. These methodologies are one level higher than databases and knowledge bases. As such, techniques and algorithms used for data mining may not be suitable.

1.4. Problem Definition

State-of-the-art technological methods have inundated private and public sector databases with terabytes of potential information waiting to be tapped. Systems that are built on this cornucopia of information must be able to comprehend and leverage this data using concise and efficient models. Many areas that computing is applied to, can be represented by domain models. Consequently, the models that represent a given domain must be able to leverage commonalities in order to achieve a larger symbiotic design. One of the methodologies that suggest an approach to accomplish this is that of Corpus Based Knowledge Representation [CBKR].

A single domain can be represented in many ways that model multiple perspectives. The bigger problems then are: (1) how to make the representation of the domain more complete as concerns the knowledge modeled by the representation, and (2) how to organize the knowledge as best as possible so that it can be used for various applications. Lateral to traditional knowledge representation methods is the Corpus Based Knowledge Representation methodology [Halevy & Madhavan, 2003]. The idea of corpus based representation is that a large domain is represented using various schemas that model its many perspectives. When these representational components are brought together as a library, they form a corpus that contains more information about a domain than any one representation. The advantage is that one does not have to design a single, extensive and complete logical ontology of the domain. In CBKR, mappings are established between representations of a domain using statistics applied to the corpus of schema representations. The underlying assumption is that patterns found in the corpus can be used for a limited number of applications like query answering and searching. However, the corpus of the CBKR methodology can be fine tuned. That is, the categorization and organization of the schema components in the corpus can be improved by identifying and extracting more associations between them. This refined corpus can then be applicable to a larger class of applications based on the analysis of the corpus, such as hypothesis testing and prediction in context.

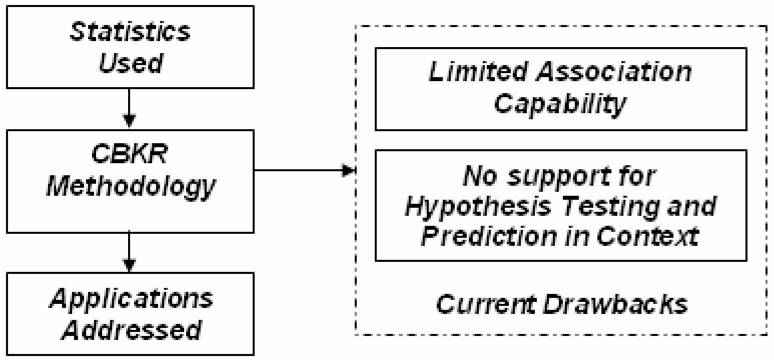

Statistics Used
CBKR Methodology
Applications Addressed
Limited Association Capability
No support for Hypothesis Testing and Prediction in Context
Current Drawbacks

Figure 1.2.: Problems with the CBKR Methodology

Figure 1.2 is a graphical representation of the research problem in this thesis. It shows the CBKR methodology that uses the statistics defined for CBKR to address a limited class of applications. The methodology has two identified drawbacks: (1) limited association capability and (2) no support for hypothesis testing and prediction in context. The lack of association capability comes from the fact that the methodology does not have criteria in place to extract substantial associations in the corpus. Applications that require the support of analysis of the corpus, and inference based on these analyses, such as hypothesis testing and prediction in context, are not supported. The CBKR methodology is limited to applications based on simple query-processing. The researcher presents a solution to overcome these drawbacks through this work.

1.5. Research Contributions

Any representation of a domain is restricted by factors such as the expertise of the designer, the mechanisms suitable for design, the knowledge that can be contained, and so on. It is almost impossible to build one complete, logical, homogeneous representation that models a large domain completely. [Madhavan et al., 2002] However, many representation schemas put together may create a more complete, more rigorous model of the domain.

To provide a solution to the drawbacks identified in section 1.4, this researcher proposes the Corpus Based Knowledge Representation Plus [CBKR+] framework. The framework aims to increase the expressiveness of the corpus-based approach by identifying and incorporating association criteria to allow the support of new forms of analyses related to hypothesis testing and prediction in context that can be applied to domains such as criminology. This framework reveals a variety of properties of the domain, and helps to improve the completeness and expressiveness of the domain's representation. It does not attempt deep semantic understanding as in traditional knowledge representation schemes. The new framework leads to the following contributions of this thesis:

- Improves the expressiveness of the representation by identifying associations in the corpus of schema representations using a suggested Set of Criteria for Association,
- Interprets a traditional domain categorization model, called Brian Gaines' Domains of Reasoning model [Gaines, 1987], to define a new domain categorization model called the Basis for Categorization model,
- Incorporates the Basis for Categorization model to (1) facilitate a first level categorization of the schema components in the corpus into possible associations and association makers, and (2) define the Set of Criteria for Association, so that the set covers all types of associations and association agents,
- Defines analysis mechanisms to identify and extract further associations in the corpus through the suggested Set of Criteria for Association, so as to perform hypothesis testing and prediction in context,
- Demonstrates the framework by (1) showing how the new approach can enhance the analyses that can be performed by the original methodology, and (2) applying the framework to the physical world domain of Criminology.

The ability to test hypotheses and predict future possibilities using the suggested CBKR+ framework is based on emergent behaviour that results from observing a sufficient number of regularities in the domain. While CBKR is a competent methodology for representing and using knowledge, the proposed CBKR+ framework enhances its competence in an elegant and easy to comprehend manner by improving the organization of the components of knowledge contained in the domain, and by widening the spectrum of applications that the methodology can address.

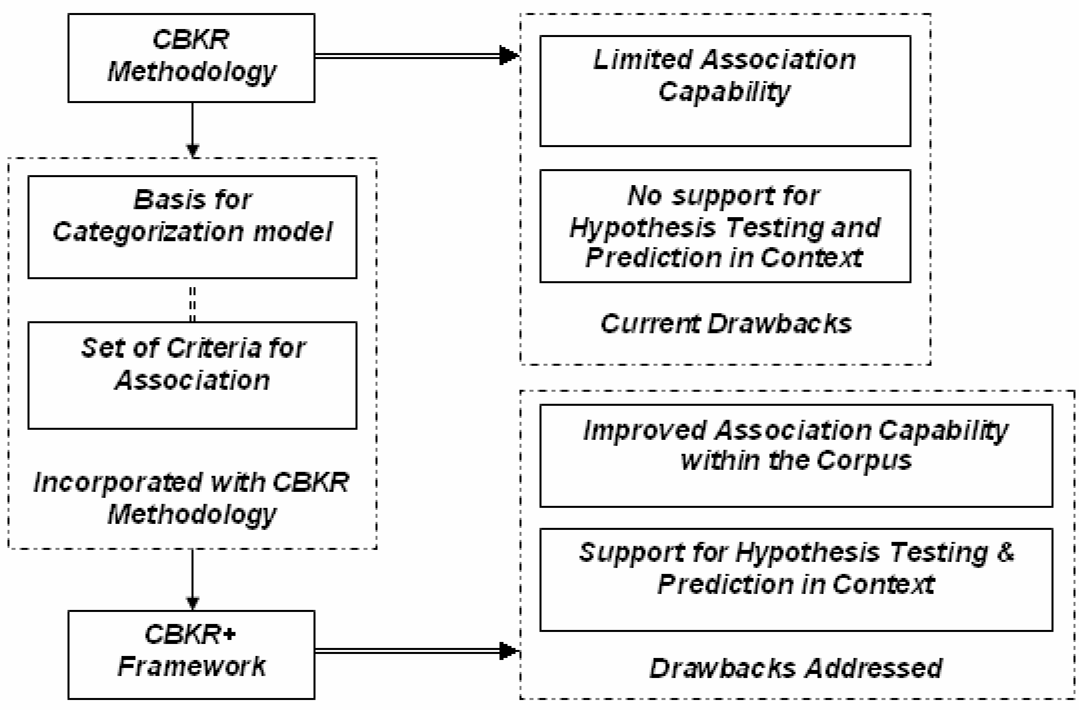

CBKR Methodology
Limited Association Capability
No support for Hypothesis Testing and Prediction in Context
Current Drawbacks
Basis for Categorization model
Set of Criteria for Association
Incorporated with CBKR Methodology
Improved Association Capability within the Corpus
Support for Hypothesis Testing & Prediction in Context
Drawbacks Addressed
CBKR+ Framework

Figure 1.3.: Evolution of the CBKR+ Framework

Figure 1.3 shows the CBKR+ framework. The CBKR methodology is faced with the current drawbacks of (1) limited association capability, and (2) no support for hypothesis testing and prediction in context. The Basis of Categorization model and the Set of Criteria for Association are incorporated with CBKR to form the CBKR+ framework. Via the framework, (1) fine-tuning of the corpus and improvement of the expressiveness of the domain representation by identification and extraction of associations between the components of the corpus shall be done, and (2) additional applications – hypothesis testing & prediction in context – based on analysis of the corpus will be addressed.

This research is about fine-tuning the corpus of collated representational components by using a mechanism of identifying and extracting associations for organizing the components in the corpus. The mechanism helps to refine the corpus by segregating and categorizing the representational components of the schemas that model the domain knowledge. While there are other forms and models of identifying associations between the components, the main asset of the proposed framework is that it is simple.

A methodology is defined by Merriam Webster as “(1) a body of methods, rules, and postulates employed by a discipline: a particular procedure, or (2) the analysis of the principles or procedures of inquiry in a particular field.” [Merriam-Webster, 2006a] Merriam Webster defines a framework as “a basic conceptional structure” [Merriam-Webster, 2006b]. According to Zachman, a framework is “a logical structure for classifying and organizing complex information.” [Zachman, 1997] In this research, CBKR is called a methodology because it is a set of rules and procedures applied to a corpus of schemas that represent a domain. CBKR+ is called a framework because it incorporates the Basis of Categorization model and the Set of Criteria for Association with CBKR in order to enhance the technique. CBKR+ is a basic conceptional structure for classifying and organizing the components in the corpus, and is thus termed as a framework.

1.6. Thesis Organization

The thesis is organized as follows: Chapter 2 reviews traditional knowledge representation systems, and serves to establish that most of these techniques are complex and cannot fully represent a domain. It discusses the background of the CBKR methodology, and research related to the Domains of Reasoning model. Chapter 3 is a detailed explanation of the proposed CBKR+ framework, including the Basis of Categorization model and the Set of Criteria for Association. This chapter also discusses the CBKR+ framework on the basis of Torsun’s criteria introduced in section 1.2. Chapter 4 is a demonstration of the framework with an example from, and as applied to the domain of Criminology. Chapter 5 discusses some of the drawbacks and biases of the proposed model, presents the conclusions arrived at in this research, and finally, suggests some future directions for this work.

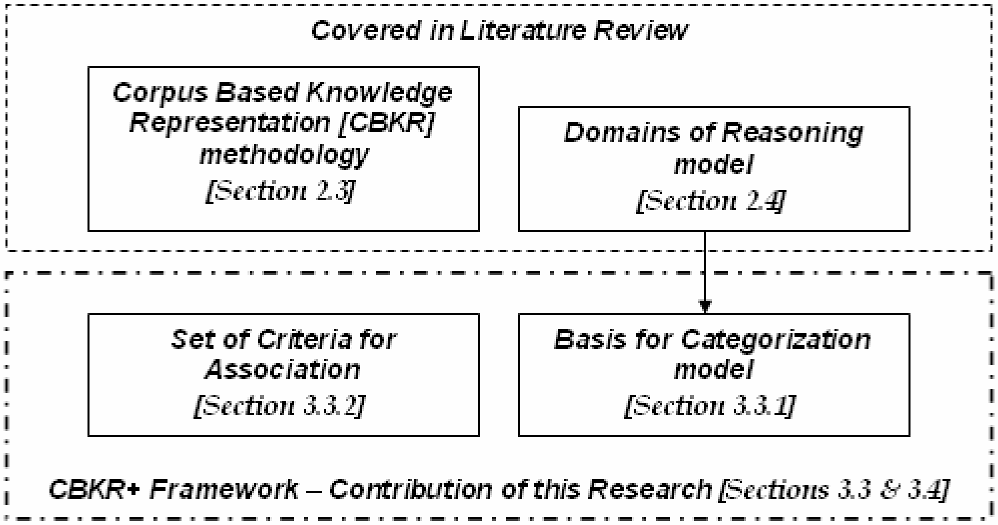

Covered in Literature Review
Corpus Based Knowledge Representation [CBKR] methodology [Section 2.3]
Domains of Reasoning model [Section 2.4]
Set of Criteria for Association [Section 3.3.2]
Basis for Categorization model [Section 3.3.1]
CBKR+ Framework – Contribution of this Research [Sections 3.3 & 3.4]

Figure 1.4.: Thesis Organization

Figure 1.4 indicates the sections that cover a particular part of the CBKR methodology and the CBKR+ framework.

1.7. Chapter Summary

This chapter introduced the problems with the implementation of any knowledge representation system. This was followed by a short discussion on building and evaluating any knowledge representation. The motivation for conducting the research in this thesis was discussed next, followed by the problem definition for this research, and the proposed research contribution of this thesis. The chapter also indicated how the thesis is organized.

Chapter 2 – Background & Related Research

When you reach the end of what you should know, you will be at
the beginning of what you should sense.
~ Kahlil Gibran (1883-1931), Philosopher and Poet

2.0. Chapter Introduction

A single domain can be viewed in many different ways. That is, multiple representations of different aspects of a domain will yield a more comprehensive representation of that domain when viewed collectively. It may be possible to build a library of all the different representational components of a single domain. Using this library instead of a single representation schema will make the representation less brittle in terms of the modeled knowledge of the domain. To make the system as efficient and flexible as possible, care must be taken that this library not be built in any arbitrary manner. There must be some mechanism, some form of underlying reasoning, to segregate, categorize and catalogue the representational components, as is done in any good library. The representational components of the domain can then be categorized on the basis of these sub-divisions. This is a wholly new approach to knowledge representation. It is not one that lies vis-à-vis classical schemes, but one that complements them, and adds to the repertoire of methods to represent knowledge of a given domain and use it for various purposes.

This new approach is the Corpus Based Knowledge Representation methodology [Halevy & Madhavan, 2003]. CBKR matches different representational schemas of a given domain by creating associations, or mapping between the various representational components of the schemas. The various schemas of the domain contributed by various sources are considered to be a library, or a corpus, of representational components. The process of matching and mapping, if done well, creates a more symbiotic representation of the domain, wherein domain knowledge is more comprehensively represented, and can be leveraged to address various applications more efficiently.

Since CBKR is built on the concepts of mapping and matching, the chapter starts with an explanation of matching concepts such as, the similarity measures needed to perform matching in a domain and the characteristics of schema matchers. This is followed by a detailed description of mapping between representations of domain models and some methodologies to do the same. This discussion is the foundation of the CBKR methodology. Then, the chapter focuses on a detailed description and analysis of the Corpus Based Knowledge Representation methodology itself. This analysis is followed by a comprehensive description of the Domains of Reasoning [Gaines, 1987] model. The last part of this chapter briefly discusses using this model with CBKR to indicate how the two will be related in this work.

2.1. Traditional Knowledge Representation

There are many views on knowledge, leading to various types of representations, each with its own drawbacks and level of complexity. Knowledge representation is complex and inherently unable to completely represent any domain. Among its many other implications and connotations, "knowledge is a relation between a knower …, and a proposition that may be an idea expressed by a simple declarative sentence … ."[Page 2, Levesque & Lakemeyer, 2000] A representation can be explained as a relationship between any two domains where one domain stands for the second domain. [Levesque & Lakemeyer, 2000]. Brian Smith's Knowledge Representation Hypothesis indicates that creating a knowledge representation system is like building a system using symbolic representations. This system is knowledge-based, where the symbolic representation is its knowledge base. The knowledge representation has an epistemological level and a heuristic level. [Sowa, 2000, Smith, 1982] The significance of knowledge representation in computational and socio-technical systems can be described well in terms of the distinct task-specific roles that it plays as described by Davis, Shrobe and Szolovits. The observations that every role suggests indicate that any knowledge representation faces some fundamental barriers that must be overcome to make efficient use of the representation. [Davis et al., 1993, Brachman & Levesque, 1985 as in Davis et al., 1993, Davis, 1991, Davis & Shrobe, 1983]

A knowledge representation deals with (1) the organization and processing of the knowledge of the domain, (2) the optimality and completeness of the data structures and algorithms used by the various intelligent agents that facilitate this organization and processing, and (3) the efficiency of the types of reasoning that can be done with the knowledge. [Partridge, 1996 as in Pesonen, 2002] Complexity of knowledge representation arises due to the limited ability of mathematics to represent and leverage reasoning that makes sense out of chaos in the domain [Papadimitriou, 1996, Brewka, 1991]. The science of knowledge representation faces problems that arise from the difficulty of stating abstract theories in a computational context to match those of the physical world. [Sowa & Zachmann, 1992] Given the complexity of applying a fixed set of knowledge representation formalisms across many domains, Rasmequan has explained the many views that knowledge representation can take. These views describe knowledge representation in terms of its philosophical foundations, its psychological associations, its use in real world business operations, and its computational characteristics. [Rasmequan, 2001, Russell, 1972, Reid, 1961, Polanyi, 1983, Rumelhart & Norman, 1988, Baets, 1998, and Way, 1994 as in Rasmequan, 2001, Markman, 1999].

Information is known to be categorized as temporal or spatial, and further as behavioural or structural, as well as hybrids of the four, depending on its origin and use. Primarily, research segregates the representation of knowledge as applied to the two larger domains of information type – temporal and spatial. [Bichindaritz & Conlon, 1996, Mohan & Kashyap, 1988] The working of a domain can be described as the domain consisting of symbols with semantics and processes that manipulate those symbols, resulting in the creation of new symbols. [Pesonen, 2002, Pinker, 1997 and Clark, 1993 as in Pesonen, 2002]. Representation schemes can be classified as declarative and procedural, where declarative schemes can be subdivided into logical and semantic [or network] ones. Many schemes combine features from more than one category in order to represent the needed domain. The means of implementation have evolved over the years from foundational methods such as logic, production systems and frames and scripts to artificial neural nets and genetic algorithms that are based in these foundational methods. [Mylopolous, 1980, Minsky, 1975, Schank & Abelson, 1977, Torsun, 1995, Anderson, 1995, Thagard, 1996, Partridge, 1996 and Hayes, 1985b as in Pesonen, 2002] These traditional schemes serve to demonstrate that representing the knowledge of a given domain and leveraging it for use in applications can be a costly, complex, and tedious challenge. It is also observed in the above research that one of the bigger problems with knowledge representation is that knowledge is brittle at the edges of a given domain. Brittleness of knowledge is meant to indicate that the knowledge used to represent any given domain is inherently incomplete; it is not possible to represent a given domain in its entirety as there will always be some information that is missing or that is not acquirable.

2.2. Matching and Mapping Techniques underlying CBKR

Corpus Based Knowledge Representation has its foundations in the matching and mapping of schemas where different representational schemas of a given domain are matched by mapping the elements within those schemas to one another. [Halevy & Madhavan, 2003] The following sections discuss the concepts of matching and mapping central to CBKR.

2.2.1. Schema Matching

A schema is a representation, in a given manner, of the knowledge of a given aspect and perspective of the domain, complete with facts, and conditions. For this work, a schema consists of a set of elements related on the basis of their attributes and organizational properties. A mapping results when a matching operation between two schemas results in certain elements that indicate that some components of one schema are related to those of the other. Since schemas do not completely capture the semantics of the data, there is a likelihood of there being several plausible mappings between two schemas, making the process of schema matching subjective in nature. Some of the methods to lessen the number of mappings are to guide the matching using an initial state of pre-determined valid mappings, a dictionary and thesaurus, a library of known mappings, user input, and user validation of the result, among other techniques. Schema matching is applied to various applications, such as mapping messages between different XML formats, mapping data sources into warehouse schemas in data warehousing, and identifying points of integration between heterogeneous databases using mediators. [Madhavan et al., 2001]

A measure of similarity is needed for the matching of components between two schemas. This similarity measure must be well-defined so that there is no ambiguity to what the system means by a match, the user can decide whether the system is applicable to the situation given to matching, and the system can use special purpose techniques for the matching process with ease. The similarity measure should ideally correspond to the way similarity is intuitively viewed, by focusing solely on the semantic content of the concepts involved, not on the syntactic terms of the concepts. The system must be able to handle a number of similarity measures without contention in order to increase the system's applicability. [Doan et al., 2002]

Schema matching is an area of potential research as representational schemas for identical concepts may have structural and naming differences, the schemas may model similar but non-identical content that is expressed via different data models, every schema may model a different view of similar content of the domain, the schemas may model different perspectives of the domain, they may represent similar non-identical domains, or they may use similar terminology to have different connotations and semantic associations. [Madhavan et al., 2001] The information in the following subsections is largely a summary of the work of Madhavan et al [Madhavan et al., 2001, Madhavan et al., 2002].

2.2.1.1. Schema Matchers

Schemas are matched on the basis of mapping similar concepts and corresponding data sets using schema matchers. Schema matching is an important aspect of a wide variety of applications. Therefore, the matcher must be generic in order to be cross-application compatible without compromising the robustness of the solution. The durability and sturdiness of the solution depends upon its abilities of (1) incorporating knowledge learned from previous cases to solve new cases, (2) modifying knowledge as the domain is understood better, and (3) handling multiple types of knowledge in order to maximize the matching accuracy. [Doan et al., 2001] There are different schema matchers that can be used depending on the applications being addressed. The various characteristics have been outlined below [Madhavan et al., 2001]:

- *Schema vs. instance based* – Matchers based on schemas only take into account schema information, such as names, descriptions, relationships, constraints, etc, and not instance data. To annotate the schema, matchers based on instances, on the other hand, use meta-data and statistics collected from data instances, or find correlated schema elements in a direct fashion, as in machine learning.
- *Element vs. structure granularity based* – A matcher can either compute a mapping between individual schema elements based on their attributes and properties, or

it can compare structurally sound combinations of elements in the schema that are related to one another, or have commonalities among them.

- *Linguistic based* – A linguistic matcher uses names in canonical form, arrived at by stemming and tokenization, of schema elements and other textual descriptions to match substrings of data. Using generic and domain-specific thesauri, the process of matching compares the equality of names, synonyms and hypernyms. This characteristic of schema matching is best applied to information retrieval (IR) applications that use descriptions to annotate schema elements.
- *Constraint based* – A matcher that uses constraints to perform its functions, utilizes schema constraints, such as data types, value ranges, uniqueness, its requirement, cardinalities, and intra-schema relationships such as referential integrity.
- *Matching of cardinality based*– The cardinality of the mappings between elements found by the schema matchers may be 1:1, 1:many, many:1 or many:many, depending upon the situation.
- *Auxiliary information based* – Schema matchers use different kinds of auxiliary information sources – dictionaries, thesauri, input match-mismatch information - past match information – either in conjunction or as single resources.
- *Individual vs. combinational based* – A schema matcher may work as an individual using a single matching algorithm, a combinational hybrid matcher that uses multiple criteria for matching, or a combinational multiple matcher that uses the combined results of independent match algorithms run on the two schemas.

2.2.2. Mapping between Representation Models of Domains

Mapping between representational models of domains is important for facilitating the sharing of knowledge and data between multiple representations of a single domain. When different representations of a single domain are mapped to one another, they can provide a more complete representation of that domain than a single representation alone. For example, queries answered by simple information integration systems use data sources across several databases in an enterprise, and the semantic web uses agents coordinating over multiple ontologies. It seems more feasible to develop multiple ontologies and schemas by independent entities, and coordinate them by mapping between the different models, using a set of formulae that provide the relationships between the components in the models. Most systems that use manual model mapping techniques are labor-intensive, error-prone, and make scaling-up difficult. The tools that provide support for constructing mappings are usually domain dependent and heuristics based. They work by identifying structural and naming similarities between models, or use machine learning to learn mappings. The system may require feedback from the user in order to refine the mappings. A generic robust method must be domain independent, and use these methodologies in a principled manner. For the method to function well there is a need for well-defined and explicit representations of mappings. [Madhavan et al., 2002]

2.2.2.1. Framework for Defining Representations of Mappings with Associated Semantics

Representations of mappings with associated semantics, such as a mapping language that works with fixed source and destination languages can be done using an associated framework. This framework enables mapping between models in very different representation languages without first translating the models into a common language, using a helper model in the mapping when it is not possible to map directly. This allows the representation of mappings that are incomplete or lose information. Semantics are defined in terms of instances in the domain, and interpretations map these instances to components in the models. Heterogeneity in the making of the models of a domain does not allow the representation of exactly the same perspectives of the domain, making it difficult for a mapping to map all the concepts in one model to all the concepts in another. Some properties that can determine whether a mapping from the many others generated is sufficient for a particular task and context include, the ability of the mapping to answer queries asked over a model, the inference ability of mapping formulas, and the composition of the mappings generated. For the many ontologies that do not have actual, associated instances, but have well-defined semantics, implicit interpretation can be done. The framework can be used for types of applications that need mapping, such as ontology or information integration systems, data migration and data warehousing systems. The mappings generated show the relationship between the mediated schema and the data sources schemas. For example, (1) ontology integration uses the generated mappings and merge algorithms to create a minimal ontology, (2) data integration allows queries to be answered without having to access different data sources independently, (3) information integration answers queries over a mediated schema that encapsulates only those aspects of the domain relevant to the query and the application, (4) data migration takes knowledge from an external source and merges it with some other data that exists in a different schema or ontology with the least loss of data possible, and (5) data warehousing uses a large population of data sources for applications to lessen the possibility of erroneous answers. [Madhavan et al., 2002]

2.2.2.2. Building Mapping Tools

The use of tools to facilitate the automatic creation of mappings by proposing possibilities can speed up the mapping process. Among the many methods for building these tools is one that uses a wide range of domain-independent, language-dependent and application-dependent heuristics based on structure or naming to generate mappings. In another method, the system learns mappings that are manually provided as examples for a learning algorithm. This method can isolate general patterns and propose subsequent mappings based on those patterns. The potential drawbacks of mapping tools are that mapping between models have to be specified in a host of different representation languages, and that certain concepts in one model may not have corresponding ones in the other since not all concepts in one model exist directly in the other. Some desirable properties for tools that facilitate representation of mappings include having clear semantics, having the ability to accommodate incompleteness, and allowing heterogeneity of representation models and languages. The properties of a mapping usually decide whether a mapping suits a given task, and whether it is adequate enough for it. These properties are query answerability, mapping inference and mapping composition. Query answerability is a formalization of partial or incomplete mappings that have lost some information due to the mapping, and yet can be applied for some tasks. Mapping inference determines if the mappings are equivalent, and mapping composition assists the generation of mappings between models through intermediate models that relate them. [Madhavan et al., 2002]

2.2.2.3. Some Issues with Mapping Techniques

When dealing with mappings in a logical reasoning context, it is imperative to also consider inaccurate mappings, the uncertainty about mappings, and the means of handling them. The absence of precise mapping in many contexts creates inaccuracy, while the generation of mappings involving the combination of different heuristics and learned hypotheses gives rise to uncertainty. The choice of the best mapping depends on the application of heuristics in the cases where no perfect mappings exist. Mappings can be inaccurate due to the mapping language being restricted in its ability to express more accurate mappings, or the concepts in the two models not being matched precisely. In the absence of any accurate, one-to-one correspondence mappings, heuristics may be used to select the best one from the viable mappings. These heuristics include restrictions of the language that expresses the mapping to prune the search space of possible mappings. Some other heuristics or mapping formulae are generated by applying a set of matching rules, or by evaluating some applicable similarity measures that compare and select from the set of all possible associations. Heuristics may use syntactic information [names of concepts, nesting relations between concepts], semantic information [inter-relationship between concepts, the types of the concepts, or the labeled-graph structure of the models], data instances belonging to input models [estimation of the probability of possible correspondences], and features that efficiently capture user interaction. Other than the context and the accuracy, the choice of mapping is also dependent on the cost of applying the mapping to data, as in data management applications that prefer efficient rather than exact query executions. [Madhavan et al., 2002]

2.2.3. The Learning Source Descriptions [LSD] Technique

Using a single mediated schema, data integration systems offer access through a uniform interface to multiple data sources of disparate schemas. Queries are posed against a mediated schema, which is a virtual schema that encapsulates the domain's salient aspects. Such systems employ manual construction of semantic mappings between the source schemas and the mediated schema, which can be a drawback due to the labor involved. Schema matching has been pursued using rule based and learner based approaches. Rule based approaches only use hard coded schema information such as names, structures and domain types of schema elements to match schemas. Learner based approaches use properties such as field specifications and statistics of data content to match schema elements, or they compare those elements that are known to be similar in the given schemas. The approach of attribute matching, where a text string consisting of all meta-data on an attribute is associated with the attribute, and can be used to match attributes based on the similarity of the text strings. The information in the following paragraphs is a summary of the work primarily done by Doan et al [Doan et al., 2001].

Work related to the LSD system also centers around the concept of value correspondences that specify functional relationships among related elements. The LSD approach is extensible in that it uses both schema and data information automatically. The LSD system also makes use of data information such as word frequencies and field formats. The LSD system uses attribute matching in its base learner, and combines the base learner’s output with the meta learner. Value correspondences and the LSD system work in conjunction where the mappings produced by the LSD system are used as an input for the value correspondence.

Learning Source Descriptions [LSD] is the basis for the development of Corpus Based Knowledge Representation, which is the focal basis of this work and is discussed further in section 2.3. The semi-automatic LSD technique is based on machine learning techniques. It uses the semantic mappings provided by the user for a small set of data sources along with the sources to train a set of learners. Each of the learners uses a different type of information found in the source schemas and the data contained within the schemas. Then, the LSD technique uses a meta-learner to find semantic mappings for a new data source by combining the predictions of the learners. The LSD technique can use domain constraints as an additional source of knowledge, as its architecture is flexible enough to allow additional learners to use new types of information. The data integration system uses a set of semantic mappings between the mediated schema and the local schemas of the data sources to reformulate a user query into a set of queries on the data sources, with the help of wrapper programs attached to each data source, which handle the data formatting transformations between the local and the integrated system data model. All this work does not give any specific basis for association forming methodologies that are employed to do the mappings.

The development of tools for finding semantic mappings is important to achieve data integration even though the process cannot be completely automated. This is done under the supposition that the LSD technique should be able to glean significant information from the manual mappings of a small set of data sources to the mediated schema to propose mappings for ensuing data sources. The LSD technique uses multi-strategy learning to discover semantic mappings, and learn from both schema and data related features. It is easily extensible to additional learners, where machine-learning techniques utilize integrity constraints and user feedback to increase the accuracy of the mappings. The input to the matcher is structured data, so that unstructured data does not have to be converted to some structured form. [Doan et al., 2001]

The working of the LSD system, essentially consisting of base learners, a meta-learner, a prediction converter, and a constraint handler, is divided into the two phases of training and matching. In the training phase, the LSD technique asks the user to manually specify the mappings for several resources, it extracts some data from each source, and creates training examples for the base learners from the extracted data, where different learners need different sets of training examples, and it trains the meta-learner, all in consecutive order. The matching phase uses the trained learners to match new source schemas, and works in the following steps: the extraction of some data from the source, and the creation of a set of associated structured elements for each source-schema element, the application of the base learners to the structured elements, followed by the combination of the learner's predictions using the meta learner and the prediction converter, and the generation of 1:1 mappings for the target schema, using the predictions and the available hard or soft domain constraints with the constraint handler. Hard domain constraints are those that cannot be violated at all, and soft domain constraints are those that all allow minimal violation. User feedback either accepts the mappings, or has the constraint handler come up with a new set. [Doan et al., 2001]

2.2.4. The Mapping Knowledge Base [MKB] Technique

Among the other methods of schema matching are those that use previously mined matching results. They mine knowledge from a corpus of known schemas and their mappings, and use the mined knowledge for matching new schemas. One such methodology is that of the Mapping Knowledge Base, or MKB, that mines and retains the knowledge from previous matching operations, and uses its own methods to apply this knowledge to new matching operations. The MKB approach is based on the observation that the various tasks that accomplish matching many times are at least partially repetitive, and therefore capable of generating patterns. Every matching task is fully, partially or not at all automated, and every new mapping between the schemas creates a possibility for the extraction of generic knowledge that can be generally applied in other related tasks. The MKB methodology is based on the construction of a corpus of schemas and knowledge such as, known mappings and results from earlier matching tasks. This corpus is called the Mapping Knowledge Base (MKB). The two parts of an MKB are schemas and mappings. Schema knowledge may be names of elements, relationships, types of elements, and descriptions of elements in the schema. Mapping information includes every mapping that is registered as a set of correspondences or matches between pairs of elements of the two schemas that are under consideration. Schemas and mappings, are continuously added to it. The components of the MKB approach include a data structure that retains knowledge about the domain of the two schemas that need to be matched, and a set of techniques that attempt to approximate the perfect similarity measure. The similarity measures are created using names of elements in the schemas, various text descriptions of the domain, data instances in the domain, data types of the elements, and the structure of groups of elements in the domain. The MKB does not start off as a well designed universal schema that encapsulates all the aspects of a domain. It evolves as a data structure that captures and retains matches between the schemas in the domain and the knowledge that it receives. This section is are a summary of the work by Madhavan et al. [Madhavan et al., 2003a]

Most matching algorithms prior to the MKB technique use linguistic similarities among names in the schema as well as structural similarities of the schema [Rahm and Bernstein, 2001 as in Madhavan et al., 2003a]. However, in these methods the prompts and indicators of similarity have been taken only from the two schemas being matched. However, the MKB approach collects this information from sets of example schemas and matches to apply past experience to new matching tasks. The LSD system [Doan et al., 2001 as in [Madhavan et al., 2003a] also exploits previous matching tasks, but it does so in a very restricted context as it maps multiple data sources to one mediated schema. In the LSD approach, trained classifiers recognize the fixed set of different elements in a single schema, and then apply themselves to new schema elements that have to be mapped again to the same mediated schema. Since the MKB approach uses previous mappings for matching any pair of schemas, general purpose knowledge must be retained in addition to knowledge regarding a particular schema.

The MKB technique does not compare the information in the two schemas that are to be matched, but relies instead on accumulated knowledge from previous validated matches. The main problem in the MKB methodology is the semantic heterogeneity of the data sharing system, as in a federated database [Sheth and Larson, 1990 as in Madhavan et al., 2003a], a data integration system [Wiederhold, 1992 as in Madhavan et al., 2003a], a message passing system [BEA, 2003 as in Madhavan et al., 2003a], a web service, or a peer-data management system [Halevy et al., 2003 as in Madhavan et al., 2003a]. The problem of semantic heterogeneity arises due to the independent design of the data sources involved that results in the design and use of different schemas to suit the individual data source. The MKB approach looks at schema matching from the point of view of determining a semantic mapping of associations and expressions that specify how the data in one source of a particular schema corresponds to the data in the other source of another schema in order to obtain a meaningful interoperation of similar elements in different schemas. There can be problems with scalability with the current technique as the MKB learns a set of models for every element of every schema that it registers. Elements in the MKB can be merged and pruned based on their predictions made for different matching tasks. The learning by the system is done offline, and therefore does not present much concern about the system's computational efficiency. But, as the MKB gets larger, interactive schema matching may become increasing difficult. As yet, matching is not domain independent, in that a MKB trained on one domain cannot be easily adapted to match schemas in completely different domains.

Another related system is the COMA system that also makes use of stored mappings [Do and Rahm, 2002 as in Madhavan et al., 2003a]. For two schemas that are to be matched, COMA searches for a schema in its reuse library for which there are stored matches between the library schema and the first schema and the library schema and the second schema, and uses the stored results to produce a new match. This technique is limited in the sense that the mappings of the two schemas under consideration to the library schema have to already exist, which can be an uncommon occurrence. Another approach suggested by Berlin and Motro [Berlin and Motro, 2002 as in Madhavan et al., 2003a] is similar to the MKB approach. This system maintains an element dictionary that is incrementally added to. It is very simplistic as it uses only data examples as prompts and indicators for matching. [Madhavan et al., 2003a]

2.3. Corpus Based Knowledge Representation [CBKR]

Lateral to traditional knowledge representation methods is the Corpus Based Knowledge Representation [CBKR] methodology. The idea of corpus based representation is that a large domain can be represented using various schemas to form a corpus or library of schemas. This corpus is able to represent the knowledge of the domain more completely than any one single schema representation. The effective analysis of any large corpus is important to successful Information Retrieval (IR) and Natural Language Processing (NLP). [Halevy & Madhavan, 2003] Corpus Based Knowledge Representation comes from the belief that a large corpus of the representational components of domain models can be used for representing knowledge effectively. The scope of this research is limited to a corpus of representational components of a single domain. The biggest advantage of corpus based representation is that it does not need the difficult design of a single comprehensive ontology that must be adhered to by anyone who wishes to contribute to the corpus. Multiple ways of representing the same information, and evolving the represented information into a larger symbiotic whole, creates opportunities for exploring problems related to the heterogeneity of representation. The concept of Corpus Based Knowledge Representation is directly built over a framework that defines representations of mappings with associated semantics, such as a mapping language that works with fixed source and destination languages.

2.3.1. Knowledge Bases in a Corpus based Context

Query answering, learning and diagnosis can be done using knowledge represented declaratively in a knowledge base (KB) that is enhanced with association forming mechanisms. The researchers of CBKR claim that most domains in the physical world are complex. The representation of a large domain therefore, is costly and complex in terms of building the knowledge base that supports the representational system. The other reasons that make the construction of a knowledge base a highly labor intensive process are [Halevy & Madhavan, 2003]:

- Acquisition of domain knowledge and its expression in formalisms.
- Creation of the knowledge base by many experts and knowledge engineers in the interest of having the knowledge base form a comprehensive whole.
- Alleviation of the brittleness of the knowledge base in terms of the knowledge contained, as only that knowledge that has been anticipated in advance is included.

These problems have been significantly researched, but eliminating all of them in order to make a perfect knowledge representation system still poses a fundamental challenge. Corpus based representation analyzes the properties of a large corpus of knowledge components that include, but are not limited to, individual knowledge bases and queries written over them, database schemas, data instances of the schemas, queries written over the databases, and any meta-data associated with the above. The corpus is constructed using independent contributions from domain experts and knowledge engineers, pre-empting the need for a thorough and meticulous ontological design. The Corpus Based Knowledge Representation methodology that comes from schema and ontology matching [Doan et al., 2001, Doan et al., 2002 as in Halevy & Madhavan, 2003], is based on the perception that the large corpus will hold patterns or associations in the representational components that can be used for many applications that have knowledge intensive processes. For this methodology, matching and mapping are related such that the matching problem is to find associative mappings between the representational components of the same domain. The corpus based methodology presumes that that the construction of detailed knowledge bases for matching is not cost effective. Therefore, this method, which is not a replacement for traditional knowledge representation, bases its approach on analyzing the differences in the representations of the schemas. The corpus based method can be applied to those tasks that do not need intensively accurate reasoning and rigorously harmonized knowledge bases. [Halevy & Madhavan, 2003]

2.3.2. *Corpus based Matching*

The procurement of a knowledge base that contains detailed knowledge about the domain in which matching is going to be done suffers from limitations such as, (1) the cost of resources needed for creating a knowledge base that is comprehensive enough, (2) the complexity of creation, (3) the brittleness of the resulting knowledge base in terms of it being used for conducting matching only on that part of the domain that has been accounted for, and (4) the limitation of the perspectives provided by a single knowledge base. The knowledge collected comprises the different ways in which the terms of the database structures, such as relation names, attribute names and data values are used, and the variations in that usage. The corpus of representational components and validated mappings does not require the meticulous ontological design that a knowledge base does, or a consistent ordering of the contents. The corpus provides multiple perspectives of a given part of the modeled domain, and also allows a domain to be covered in many different ways – a single aspect with many perspectives, and many aspects with many perspectives. The corpus based methodology thus has a higher likelihood of providing knowledge that can be used towards the matching of schemas that are composed of representational components. [Halevy & Madhavan, 2003] A domain may be characterized by many representational components that are part of the various representational schemas that embody the domain. The schemas may be XML tree representations, SQL-like relational databases, MS Excel-like field records, simple text descriptors, and so on, each contributing its components to the corpus. The schemas may model the same aspect of the domain in dissimilar ways, or many perspectives. They may also model many aspects of the domain.

2.3.3. *Work related to CBKR*

The LSD and MKB approaches explain the relation between matching and mapping, as well as indicate the use of these concepts in CBKR. Components of representational schemas under consideration are matched using various rules and techniques. The foundations of Corpus Based Knowledge Representation lie in generic schema matching, irrespective of any particular data model or application, using mappings between schema components based on their names, data types, constraints, and schema structures, that are established using a broader set of techniques, such as the integrated use of linguistic and structural matching, and context-dependent matching of types that share similar attributes and properties. [Madhavan et al., 2001] The LSD approach has also applied to simple taxonomies of concepts in CBKR. [Doan et al., 2002] Further research focused on the ability and the advantages of using a corpus of schemas and matches to predict mappings between a pair of new schemas. Models for categorizing components in the corpus were learned using the information available in the schema and the validated matches present in the corpus. The corpus based methodology found matches that would not have been predicted by other techniques. [Madhavan et al., 2003]

2.3.4. Details of a Corpus based Representation System

The details of Figure 2.1 [Figure 1, Halevy & Madhavan, 2003], that explains the corpus based representation system have been elaborated after the diagram.

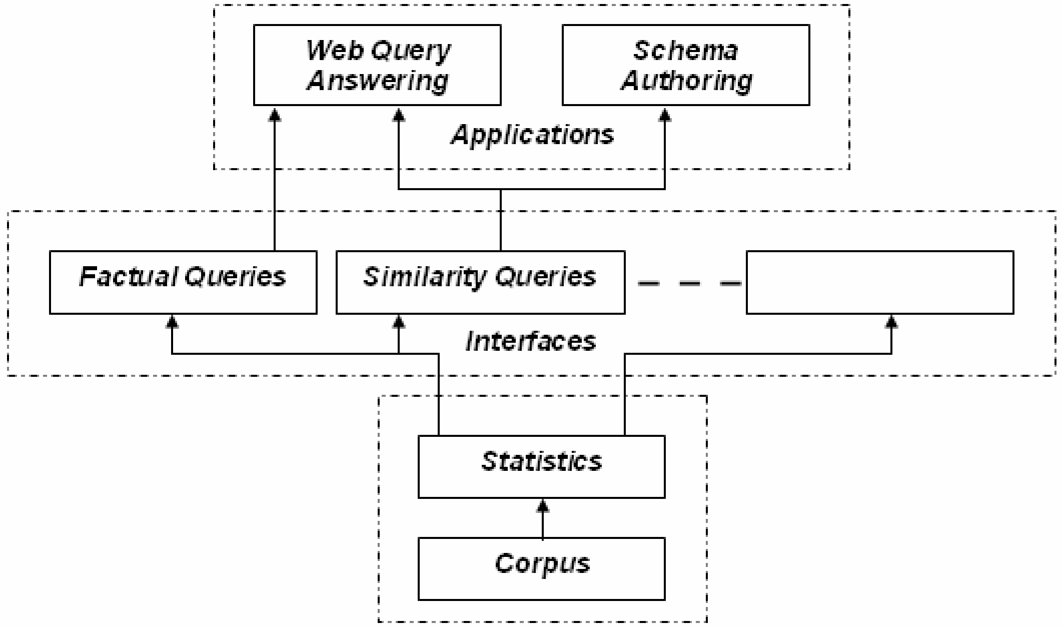


Figure 2.1.: A Corpus based Representation System

[Figure 1, Page. 1569, Halevy & Madhavan, 2003]

The figure shows the CBKR methodology, where the blank box indicates the possibility of expansion as part of future research, and has been leveraged by this work. The remaining parts are discussed below [Halevy & Madhavan, 2003]:

- *Contents of the Corpus*: The corpus itself is a collection of dissimilar structures that do not form a logically coherent universal database. The information of the domain in the corpus is manipulated using tools that work with a set of operators for manipulating the representational components of the domain and not the data elements themselves [Bernstein, 2003 as in Halevy & Madhavan, 2003]. The corpus includes, but is not limited to, information related to structured data such as, various forms of domain knowledge about the modeled perspectives of the domain, how the perspectives are segregated and classified, and so on. The representational components may comprise relational and object oriented database schemas or entity-relationship diagrams, XML data type definitions and schemas, their associated functional dependencies, and terminologies and expressions confined by an associated lexicon and grammar. The corpus also includes instance data that may represent actual elements of the domain in the form of concrete rows of tables, XML documents, components of a terminological knowledge base, and data in formats that lack a schema. Sometimes the components in one representational schema may be instance data of another, and that is acceptable as long as it is tagged likewise. The corpus may also contain direct or indirect validated mappings between the representational components of the domain, user initiated queries that can generate metadata about the domain on how data is used based on the types and frequencies of queries and the sets of answers associated with them, and other meta-data that accompany domain models that provide any other information about the domain itself, the perspective in which the domain must be viewed, and so on.
- *Statistics on the Corpus*: Analyzing a corpus and using the results towards knowledge representation comes from the area of Natural Language Processing [NLP] [Solan et al., 2005, Sowa, 2000]. The CBKR methodology makes use of statistics that can be computed over the corpus made from representational components of the same domain. [Halevy & Madhavan, 2003] These statistics are Word and Term Statistics that are run over individual words and noun/verb phrases to show the use of words in structured data, where techniques such as

word stemming, synonym tables, inter-language dictionaries, and their combinations, determine the different versions of these statistics. This class of statistics includes those like Term Usage, which shows the frequency of the term used as a relation name, attribute name, or in data, Co-occurring Schema Elements, which indicates the relation names and attributes that can usually occur in relation to a term, combinations of attribute names that exist concurrently, and mutually exclusive uses of attribute names, and similar names that show for each term, other terms that are used with similar statistical characteristics. The next category of statistics that can be run over a corpus is that of Composite Statistics. These statistics are similar to those above but can be applied to partial structures, such as sets of data instances, sets of relations with associated attribute names, and sets of relations with associated data. Statistics for partial structures must be limited with regard to space and complexity constraints. The third category of statistics is that of Statistics for Schema Elements. Such statistics are run over the particular schema to which the component belongs, and other components mapped to the schema by validated mappings. [Doan et al., 2002 as in Halevy & Madhavan, 2003]. Figure 2.2 is a representation of the set of statistics associated with the Corpus Based Knowledge Representation methodology, i.e. the block of Statistics from Figure 2.1 above:

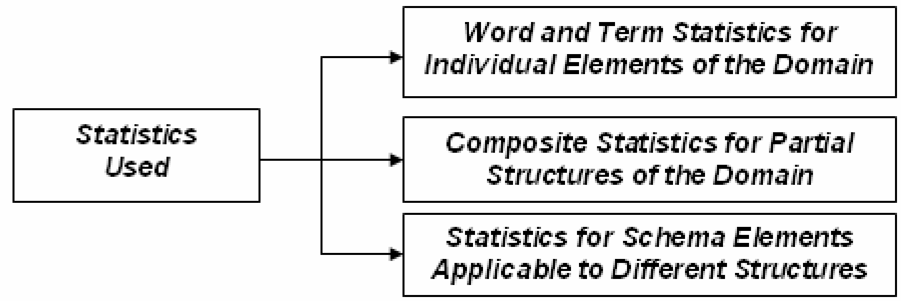

Statistics Used
Word and Term Statistics for Individual Elements of the Domain
Composite Statistics for Partial Structures of the Domain
Statistics for Schema Elements Applicable to Different Structures

Figure 2.2.: Statistics used by CBKR

– *Application Interfaces to the Corpus*: Applications that use a corpus based approach should be able to access the statistics that are run over the corpus easily, and then apply the collective results to look up queries that are put to them. The answer sets returned may often summarize sets of facts or provide descriptions of schema fragments, while also highlighting that data which cannot possibly be part of an answer to a query. Rather than responding with answers that are based on Boolean conditions on data as in logical knowledge bases, corpus based applications rank the answers according to given query conditions. The corpus based representation methodology attempts to provide interfaces to the applications it supports. Currently, these interfaces are classified on whether they support factual queries or similarity queries. When it comes to factual information queries, a complex function of constants as input returns formulas that include all of these constants, with the goal to identify the relationship between two representational components independent of the specific role they have in the schema. The returned formulas may serve as a template that creates a schema description. Additional knowledge to the factual query may help to prune the answer set, or result in a different ranking of answers. Queries based on similarity information attempt to utilize the different ways of saying similar facts, which is one of the characteristics of a corpus. A query could be an established basic fact with the corpus returning a set of similar ways of expressing the same fact.

- *Applications of the Methodology*: Dissimilar knowledge fragments and representational components of the schemas modeling that knowledge make the corpus of the CBKR methodology, over which varied statistics are computed. This method of representation presents an alternative to traditional knowledge representation for a broad class of applications that share some common properties, such as using the knowledge base to resolve ambiguity by ranking the answers and ruling out the incorrect ones [Halevy & Madhavan, 2003]. This methodology is suited to applications built on learning tasks, natural language processing and natural language interfaces, from which the method has been inspired, and, information retrieval [Popescu et al., 2003 as in Halevy & Madhavan, 2003]. In addition, classes of applications based on web search and query answering [Kwok et al., 2001, Radev et al., 2002, Sizov et al., 2003 as in Halevy & Madhavan, 2003], and creating and querying structured knowledge to close the structure chasm [Halevy et al., 2003 as in Halevy & Madhavan, 2003] stand to benefit from this methodology. CBKR is ideally not suited for applications that require intricate logical inference on homogeneous and comprehensively designed domain models, like automated vehicle control systems and deduction from judicial law.

2.3.5. Research Directions for the CBKR Methodology

The Corpus Based Knowledge Representation methodology provides several directions for future research. Some of these are the fine-tuning that can be done to the corpus in order to make the system perform better, the best means to collect a large enough corpus of interest, widening the spectrum of applications that can use corpus based representation, and the analyzes that should be performed on a corpus to make the system effective. [Halevy & Madhavan, 2003] This research focuses on improving the refinement, tuning and usability of the corpus, and making the CBKR methodology applicable to more applications.

2.4. Domains of Reasoning Model

The forming of associations in a knowledge system can be conducted using certain distinctions that define the domain. The essential concept of making a distinction is that some part of the domain is separated or marked out. The Domains of Reasoning model, explained ahead, outlines some foundational considerations for the logical organization and representation of knowledge in a given domain system [Gaines, 1987]. According to the creator of the model, it is reasonable to suppose that these essential distinctions are the generators underlying any knowledge representation. Distinctions are the operations of actions or processes in computational terms. The assumption of the model is that if the primitive process of making distinctions can be considered to motivate the knowledge processes, then it should be possible to analyze the associations between the various components of the domain in terms of these underlying distinctions. The following sections explain the Domains of Reasoning model proposed by Brian Gaines [Gaines, 1987].

Figure 2.2 shows how some essential distinctions in the domain help to establish the major categories of knowledge components within the domain [Figure 1, Gaines, 1987]. This figure has been redrawn without modifications from its source. The model intends to show that though a domain may have different semantics of truth and inference, their logical foundations are consistent within it.

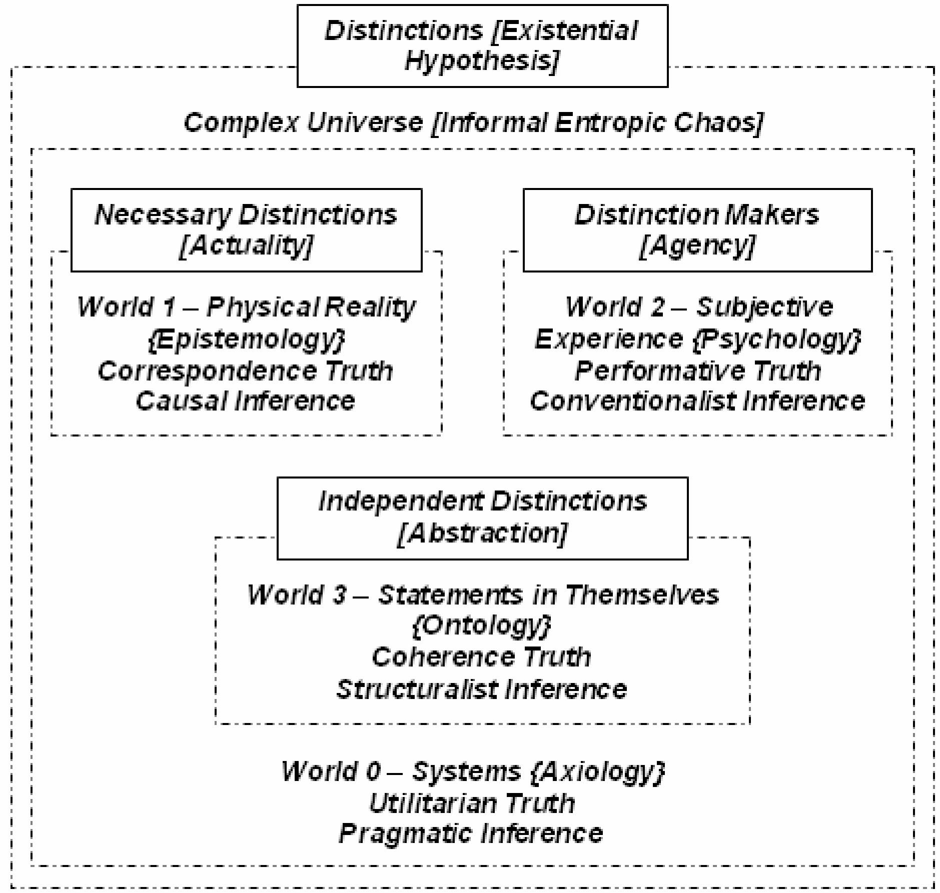

Distinctions [Existential Hypothesis]
Complex Universe [Informal Entropic Chaos]
Necessary Distinctions [Actuality]
World 1 – Physical Reality {Epistemology}
Correspondence Truth
Causal Inference
Distinction Makers [Agency]
World 2 – Subjective Experience {Psychology}
Performative Truth
Conventionalist Inference
Independent Distinctions [Abstraction]
World 3 – Statements in Themselves {Ontology}
Coherence Truth
Structuralist Inference
World 0 – Systems {Axiology}
Utilitarian Truth
Pragmatic Inference

Figure 2.3.: Worlds generated by Basic Distinctions

[Figure 1, Page. 372, Gaines, 1987]

Initially, the building material of a complex universe, including the domain to be represented, can be considered to be a kind of informal, cluttered chaos, with the only property that it is rich enough to be categorized into distinctions, or objectified and segregated into groups. The distinctions in the domain give rise to sub-domains. In World 0, the only justification of making distinctions and creating sub-domains is that it is useful to make them. Every sub-domain contains utilitarian truth and pragmatic inference, generically speaking, and is called the World of Axiology. Within this sub-domain, there will be Necessary Distinctions, Independent Distinctions, and Distinction Makers.

The sub-domain of Actuality or Necessary Distinctions contains those distinctions that are necessitated by the actual, physical domain. This World 1 is based on correspondence truth and causal inference, and is called the World of

Epistemology since the distinctions made are based on their value in modeling the physical world. As above, the sub-domain of Independent Distinctions or Abstractions contains those types of distinctions that are considered apart from their origins. This World 3 justifies the distinctions made in terms of criteria that determine that the distinctions belong together, and is one of coherence truth and structuralist inference. It may also be designated as the World of Ontology since the distinctions made exist by their own right. Distinction makers are needed to make the distinctions in a domain, giving rise to the sub-domain of Agency or Distinction Makers. This World 2 is built on subjective experience, and is based on performative truth and conventionalist inference. It is called the World of Psychology since the distinctions made are distinguished by the agent that made them. [Gaines, 1987]

2.4.1. Need for the Ability to Form Associations in CBKR

Analysis of the CBKR methodology indicates that it can be improved by enhancing it with the ability to form more associations within the components of the corpus. The action of forming associations can generally be based on the context of the domain. Forming associations may be based on analogy, where a comparison is done between various structured, abstract or established relationships, or on signs that are a surrogate for things that are abstract or not directly observable. It may be based on cause, where the relationships are based on the influence of one thing on the other, and are explained by causal inference or on testimony, which cites other claims to support ones own, or it may be based on narrative, which uses story-like constructions as support to a claim. [Norton, 2005] Knowledge is more complex than data or simple information, since knowledge is semantic and associative in nature, and as a result, knowledge maybe incomplete or incorrect as regards the domain. Methods of forming associations that can arrive at approximate answers are more useful to knowledge-based systems. This ability to accommodate more gradual correctness is advantageous to applications where approximate answers are more efficient than typically correct answers. [Utrecht, 2005]

In classical terms, the process of forming associations uses inductive, deductive and abductive inferences. The three forms share an interesting relationship. Deduction reveals essential consequences on the basis of a logical consistency between premises and conclusion that can be rationally proven.

Induction provides an observed logic between the premises and experience, which can also be proved, to derive a possible generalization. Abduction, on the other hand, not only classifies the data, but also offers a theory that attempts to justify the causal relation that exists between the facts. Induction is best used for quantitative verification, while abduction and deduction work with conceptually comprehending premises, facts and relations. When used together, the goal of induction is to gradually approximate the truth to establish a set of beliefs for further inquiry, while deduction builds logical hypotheses based on other plausible arguments, and abduction explores available data, and identifies patterns, to propose credible hypotheses based on appropriate categories. [Williams & Colomb, 2002, Hoffmann, 1997, Wirth, 2005] Therefore, in sequence of use, "abduction creates, deduction explicates, and induction verifies." [Yokasu, 2005, Page 6]

Section 2.3.5 listed fine-tuning the corpus of collated representational components in CBKR as a research focus of this work. This research proposes using a mechanism of forming associations within the components in the corpus. The mechanism will help to refine the represented knowledge as it is added to the corpus by segregating and categorizing the representational components of the schemas that model the domain knowledge. This research considers the associations formed to be macroscopic since the mechanism uses the components of the schemas that represent the domain, and not the microscopic knowledge elements contained within them. It is suggested that the integration of such a mechanism can alleviate the problems associated with fine-tuning of the corpus to some degree.

2.4.1.1. Using the Domains of Reasoning Model with CBKR

Corpus Based Knowledge Representation uses many representations of a single domain, covering same or different perspectives with similar or additive knowledge. A library of representational components is created with knowledge about the structure of the domain. This library can be leveraged for applications such as, web query answering and schema authoring, using certain statistics that are run over the corpus of representations. The organization of the corpus can be fine tuned, and the number of types of applications that can make use of CBKR can be increased. This can be done if some form of finding associations within the components of the corpus is incorporated in the creation and use of the corpus. As stated by Gaines, the distinctions made in everyday life do not have hard boundaries [Gaines, 1987]. Yet, it is possible to segregate a domain on the basis of the distinctions this model defines for a given domain. The Basis of Categorization model, introduced in the next chapter, is an instance of the Domains of Reasoning model as applied in a computational context. It is suggested that the distinctions of the Domains of Reasoning model can be applied to the CBKR methodology in the form of the Basis of Categorization model proposed by this research to perform a first-level categorization of the domain into possible associations and association makers. This categorization will facilitate the forming of associations in the corpus. The process, as well as the various models part of this work, is explained in the next chapter.

2.5. Chapter Summary

This chapter discussed the background information needed for this research. It discussed the LSD and MKB techniques of matching, essential to the process of mapping, which creates associations between the representational components in a corpus of a domain. The chapter then discussed the CBKR methodology itself in detail. This chapter also presented a description of the Domains of Reasoning model, followed by a brief indication of the use of this model with CBKR.

Chapter 3 – The CBKR+ Framework

> *There are things known and there are things unknown, and in between are the doors of perception.*
> *~ Aldous Huxley (1894-1963), Novelist and Essayist*

3.0. Chapter Introduction

In this digitized day and age, academic, corporate and government databases have to deal with massive amounts of information. The challenge of having to work with large amounts of information is creating a need for representational systems that model extensive domains with simplicity and ease of use. When a domain is modeled for computational purposes, the knowledge of the domain is modeled by the components of the representational schemas of the domain. In addition to indexing this data efficiently, the domains must be able to use competent and effective mechanisms for applications that use these schemas and the domain knowledge contained therein. It is seen from section 2.1 that traditional representational systems are based on principles of reason and intuition. However, these systems characteristically rely on time-consuming and difficult hand coding. Over the years, these systems have become denser and more complicated, and their simplicity and elegance is often a matter of concern.

This research focuses on an alternative form of knowledge representation lateral to traditional schemes called the Corpus Based Knowledge Representation methodology. There are many domains where data changes rapidly, but the essential structure of its representation does not alter significantly, such as demographic databases. One of the properties supported by applications based on these domains is that the applications are more domain schema reliant rather than domain data reliant. The components of the representational schemas of such domains may vary, but are not drastically different. [Waltz & Kasif, 1995] The CBKR methodology applies itself to these types of domains. The methodology has certain drawbacks that must be addressed in order to improve the methodology. This chapter is a discussion of the improvements suggested to the CBKR methodology in the form of the CBKR+ framework.

The CBKR methodology is designed to use a corpus of different representational schemas of a given domain towards applications such as simple query processing and information retrieval. However, the methodology lacks a method for forming extensive associations between the components of the corpus. There is immense scope for fine-tuning the corpus and improving the expressiveness and analytical capacity of the methodology. Addressing these drawbacks will make the methodology applicable to a wider class of diagnostic and probability based applications, such as hypothesis testing and prediction in context. This research incorporates a Basis of Categorization model and a Set of Criteria for Association in the CBKR methodology in an attempt to improve the CBKR approach. The Basis of Categorization model that performs a first-level segregation of the components of the corpus into possible associations and association makers is an instantiation of Brian Gaines' Domains of Reasoning model [Gaines, 1987] described in the previous chapter. The Set of Criteria for Association contributes eight criteria that identify and extract associations in the components in the first-level segregated form of the corpus. This chapter also shows that the Set of Criteria for Association is related to the Basis of Categorization model so that the criteria proposed cover the domain as comprehensively as possible by working with all possible associations and association makers. The Basis of Categorization model, the Set of Criteria for Association, and the method in which they are incorporated with CBKR and applied to a given domain form the proposed Corpus Based Knowledge Representation – Plus [CBKR+].

This chapter is organized as follows: the introduction to the chapter is followed by a brief review of the characteristics of the CBKR methodology. Then, the issue of forming associations between the components of the corpus is discussed as an enhancement to the methodology. Towards this goal, the Basis of Categorization model and the Set of Criteria for Association are explored. This is followed by an explanation of the relationship between the Basis of Categorization model and the Set of Criteria for Association. This is all put together to derive the CBKR+ framework. The chapter concludes with an appraisal of the framework using criteria suggested by Torsun [Torsun, 1995 as in Pesonen, 2002] previously explained in Chapter 1.

3.1. The CBKR Methodology – Possible Improvements

The Corpus Based Knowledge Representation methodology can be simply described as a method that builds and uses a library of domain representations and schemas and the knowledge contained within them. [Halevy & Madhavan, 2003] The intuitive assumption is that a single domain modeled from many perspectives, and in different ways, will bring together more knowledge about the domain, thereby improving the completeness of its representation. The CBKR methodology is not designed for applications based on intensive logical inference that must adhere to strictly designed domain models, and it cannot be used for applications that use multiple domains at the same time. The methodology can be applied to applications that do not need a very expressive corpus that supports enhanced analytical capability, such as web searching and query answering. The applications addressed by this methodology share the common characteristic that the knowledge base resolves ambiguity by ranking and rejection, where the candidates of an acquired result set are ranked according to some predetermined criteria, and those that fail to meet the criteria are discarded. These applications are based on the paradigms of learning tasks, natural language processing [NLP], natural language interfaces [NLI] and information retrieval [IR]. As mentioned by Halevy [Halevy & Madhavan, 2003], there are certain drawbacks of this approach. Some of these are:

- *Limited association capability*: The corpus of components acquired by collating various schemas and representations of the domain has immense scope for fine-tuning in order to increase its analytical capacity and make CBKR methodology more expressive. The CBKR methodology has limited association capability and cannot identify and extract extensive associations between the components of the corpus. This drawback can be alleviated, in some measure, by incorporating criteria for associations in the methodology.
- *No support for 'Hypothesis Testing' and 'Prediction in Context'*: Due to the above drawback, limited applications lend themselves to the CBKR methodology. It cannot address diagnostic and probability based applications that require analysis of the corpus. Such applications include hypothesis testing and prediction in context.

Figure 3.1 shows the researcher's interpretation of the CBKR methodology, and its drawbacks that will be addressed through this research.

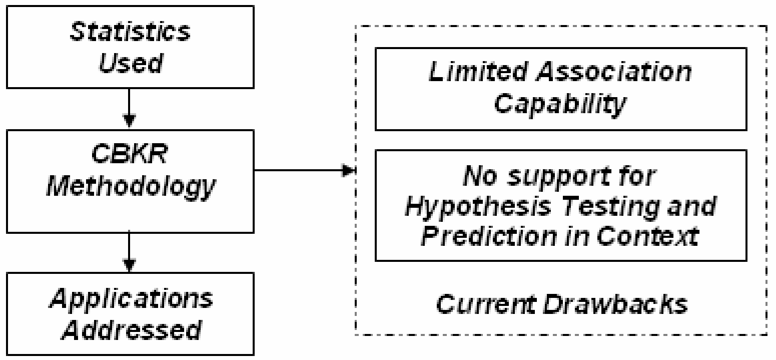

Statistics Used
CBKR Methodology
Applications Addressed
Limited Association Capability
No support for Hypothesis Testing and Prediction in Context
Current Drawbacks

Figure 3.1.: Drawbacks of the CBKR Methodology

The figure shows a snapshot view of the CBKR methodology as discussed in the previous chapter, and specifically lists the drawbacks to be addressed. The block called Statistics Used shows the set of statistics that the methodology uses. These statistics include [Doan et al., 2002]:

- Word and Term Statistics that are associated with individual words and noun or verb phrases and they indicate how a component may be used in structured data.
  - Term usage indicates how frequently a term is used as a relation, attribute, or as data itself.
  - Co-occurring schema elements indicate which relations and attributes tend to appear with different uses of a term, and what attributes seem to be together.
  - Similar names indicate terms used with similar statistical characteristics.
- Composite Statistics are the above statistics that are applied to partial structures of components within the domain, such as sets of data instances, sets of relations with their attributes between terms or sets of relations with other data, and so on. The statistics for other related structures could be estimated if statistics for certain partial structures are given.
- Statistics for Schema Elements, such as mappings for terms across structures and relation of term usage across structures, characterize the specific usages of terms in structures, and relate them to usage of terms in other structures of components. The same term, used in different structures, can have different connotations.

The block labeled 'Applications Addressed' represents those applications that efficiently use the methodology to conduct processes such as web searching and query answering.

The CBKR+ framework proposed in this research eliminates the shortcomings of the methodology as will be seen in the sections ahead. The approach of this work is that identifying associations in the corpus is not necessarily based in running symbols through complex formulas and rules, but is based in ascertaining sufficient regularities and patterns in the system and quantifying them.

3.2. Forming Associations between the Components of the Corpus

This section discusses the process of identifying and extracting associations between the components of the corpus. This is explained in the following example.

Consider a domain of crime where one representation of the domain focuses on drug trafficking. The representational schema that describes this domain is made up of components and relationships among those components. Say,

- Set of components: PERSON, DRUG, and PRISON
- Set of identified associations:
  - PERSON → DRUG: stands for the relation: component PERSON deals component DRUG
  - PERSON → PRISON: stands for the relation: component PERSON is in component PRISON

Based on observed behaviour through recorded convictions, the association DRUG → PRISON then can be inferred as a presumption that if dealing component DRUG, then will be confined to component PRISON. Figure 3.2 shows these relations. Using adequate association forming mechanisms and analyses, the association can be formed that since component PERSON deals component DRUG, and since component PERSON is in component PRISON, any PERSON dealing DRUG will be confined to PRISON. Result sets of various individuals who have been dealing various forms of coke and have been confined to various prisons can then be created for a demographic profiling of individuals having a higher likelihood of committing the crime of dealing. This can be used to set out All Points Bulletins or Advisory Notices for localities with a high concentration of such individuals.

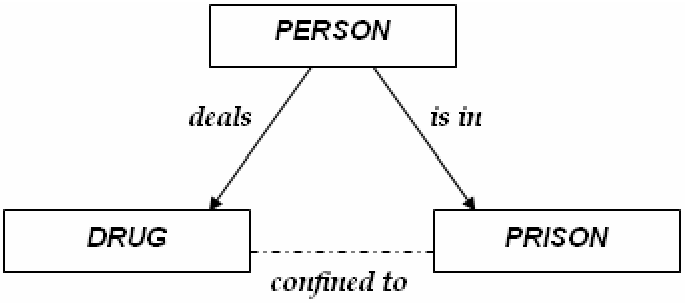


Figure 3.2.: Forming Associations between Components of a Corpus

A hypothesis can be tested against an adequate corpus using truth-values, pre-determined probabilities and confidence intervals. With regards to this example, based on the regularity of the invocation of these relationships, the probability or truth-values of hypotheses such as: ‘every person who is in prison deals coke’ and ‘every person who deals coke is in prison’ can be established. It however does not mean that every component PERSON who is in component PRISON deals component DRUG. The important factor here is that the forming of associations is based on the regularity of occurrence in the corpus. Unknown, ambiguous components, or erratic relationships between them must be dealt with different methodologies, the suggestions for which are beyond the scope of this research.

This mechanism can be extended to include a larger section of the corpus, as represented in Figure 3.3, below. By way of example, the structure of drug trafficking in the corpus of social crime is divided into partial structures such as PERPETRATOR, INSTRUMENT, and STATUS, where component PERSON is one of the components in partial structure PERPETRATOR, component DRUG is one of the components in partial structure INSTRUMENT, and component PRISON is one of the components in partial structure STATUS. If a hypothesis is found to be correct, it serves to identify more associations between the components.

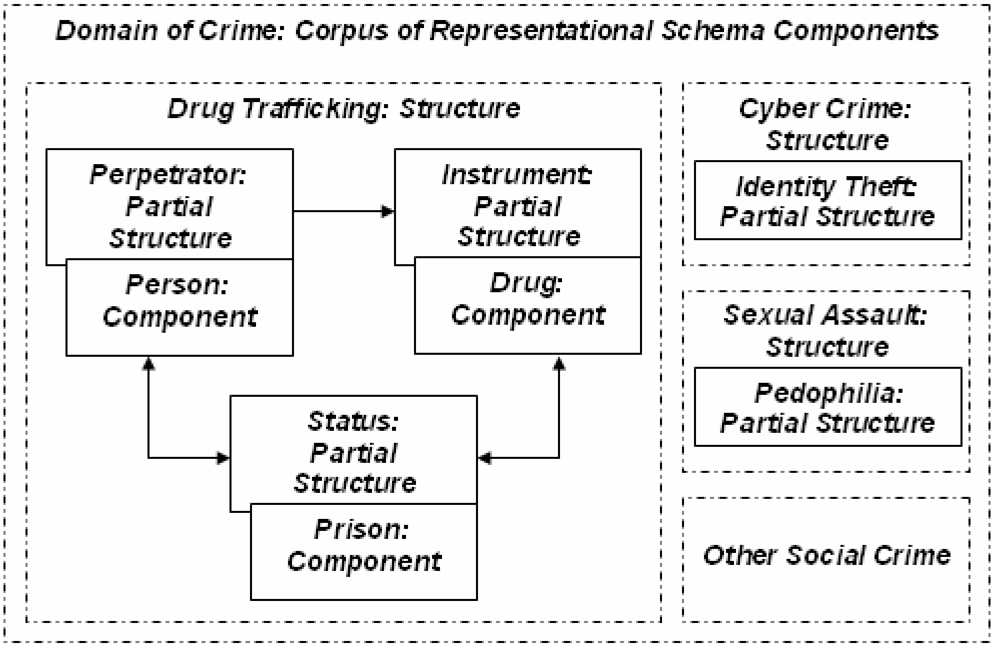

Domain of Crime: Corpus of Representational Schema Components
Drug Trafficking: Structure
Perpetrator: Partial Structure
Person: Component
Instrument: Partial Structure
Drug: Component
Status: Partial Structure
Prison: Component
Cyber Crime: Structure
Identity Theft: Partial Structure
Sexual Assault: Structure
Pedophilia: Partial Structure
Other Social Crime

Figure 3.3.: Establishing Associations between the Components in the Corpus

The corpus can be as refined as possible, but one should be aware of the trade-off between refinement and convolution in the associations extracted in the corpus. Too much refinement may make it difficult to extract clear-cut and unambiguous associations.

3.3. Forming the CBKR+ Framework

The idea underlying corpus based representation is that an extensive library of domain-model representations can be used effectively towards knowledge representation, in a manner complementary to traditional representation schemes. The advantage is that the need for a strictly homogeneous logical design of an extensive and complete representational ontology is avoided. [Doan et al., 2002, Halevy & Madhavan, 2003] A knowledge base, as in a corpus of schemas, with a mechanism for forming associations between its components can effectively address tasks such as analysis based query answering, probability based prediction, and diagnosis where a possible state of the system can be diagnosed by testing a hypothesis posed on that state.

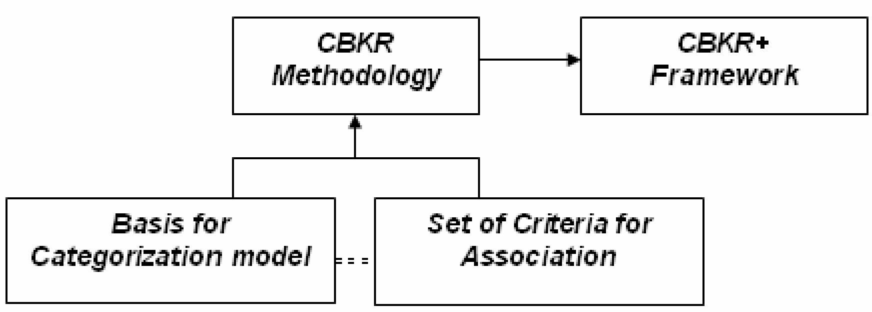

CBKR Methodology
CBKR+ Framework
Basis for Categorization model
Set of Criteria for Association

Figure 3.4.: Enhancement to the CBKR Methodology

Figure 3.4 is a high-level representation of the focus of this research. The Basis of Categorization model and the Set of Criteria for Association is incorporated with the CBKR methodology to form the CBKR+ framework. The CBKR+ framework is based on collecting a corpus of disparate fragments of knowledge in the form of schemas and representations that model the domain, identifying and extracting associations between the components of the corpus, analyzing the properties of the corpus through the associations, and using this analytical capability to make the CBKR methodology more expressive and applicable to a wider class of applications. The advantage of the CBKR methodology is that the corpus can contain a heterogeneous collection of schemas contributed by many experts, where the domain representations do not have to adhere to a strict creation method. This advantage is extended to the CBKR+ framework. One of the underlying bases of the CBKR+ framework is that if the corpus is large enough, the patterns identified by observing a sufficient number of regularities within a system in forming associations can be used for knowledge intensive tasks. [Iyengar & Bastani, 1992, Halevy & Madhavan, 2003] The corpus contains schemas that represent the structure of the domain. The components of the corpus represent a macroscopic view of the domain. As a result, this work identifies and extracts macroscopic associations. A microscopic view of the domain would be one where every entity and element within the domain, and not just its structure, would be modeled.

Once associations have been extracted, and analysis is being done, ambiguity in generating the result sets to diagnostic and probability based queries could be resolved by establishing a set of predetermined confidence levels. If the results of the hypothesis being tested fall within the confidence interval, the truth value of the hypothesis can be said to be true. Similarly, if the regularity of certain observed patterns falls within the confidence interval, a prediction in context of those patterns can be said to have a high probability of occurrence.

The method of the CBKR+ framework is sequentially co-dependent – first, the Basis of Categorization model uses existing components in the corpus and strictly validated relationships between them to create a structure for finding possible associations and association makers in the corpus. This can be called the inductive process, where specifics are leveraged to arrive to general conclusion about all of them. Then, the Set of Criteria for Associations identifies and extracts specific associations in the corpus. This can be called the deductive process since it progresses from general warrants and reasons to specific claims.

Figure 3.5 is a more detailed version of this transition, as developed in this work. The double dotted line divides the figure into two halves. The left-hand-side of the figure has been adapted from the original CBKR model [Halevy & Madhavan, 2003], which has been discussed in the previous chapter. The right-hand-side of the figure is the contribution of this research.

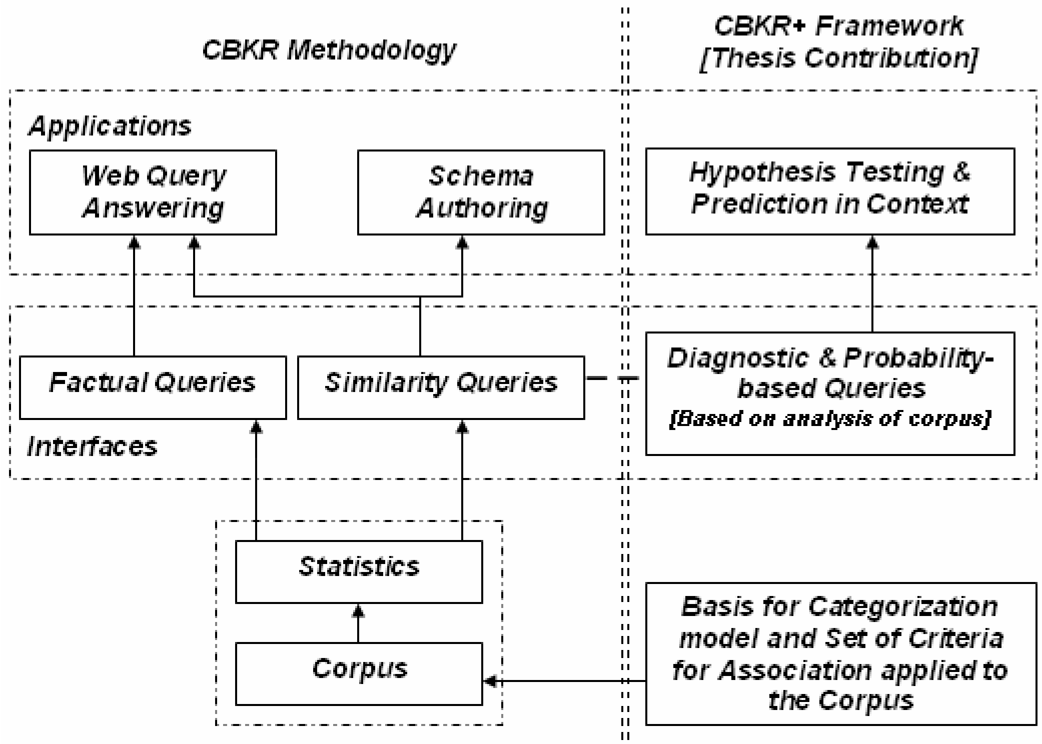

CBKR Methodology
CBKR+ Framework
[Thesis Contribution]
Applications
Web Query Answering
Schema Authoring
Hypothesis Testing & Prediction in Context
Factual Queries
Similarity Queries
Diagnostic & Probability-based Queries
[Based on analysis of corpus]
Interfaces
Statistics
Corpus
Basis for Categorization model and Set of Criteria for Association applied to the Corpus

Figure 3.5.: The CBKR Methodology enhanced to form the CBKR+ Framework

The CBKR methodology employs Statistics, such as word and term statistics, composite statistics and statistics for schema elements to study the Corpus of collated schemas. It uses Interfaces in the form of Factual Queries and Similarity Queries to apply itself to applications such as, Web Query Answering and Schema Authoring. As part of the CBKR+ Framework, the Basis of Categorization model and the Set of Criteria for Association, described later in this chapter, use Diagnostic and Probability Queries as the interface. This interface makes use of the analysis done on the corpus in the form of extracting associations between its components to address applications such as hypothesis testing and prediction in context. The domains for such applications must be more domain schema reliant rather than domain data reliant. For example, demographic databases can be modeled in many ways, but essentially contain similar content.

3.3.1. The Basis of Categorization Model

Intuitively, it can be assumed that most features within a domain are interdependent and have associations that relate them. For a represented world, the modeling of the domain and the knowledge contained therein is done via schemas. The Domains of Reasoning model [Gaines, 1987] is built on the assumption that any mechanism for finding associations in the domain can be analyzed, and then implemented if the knowledge processes in a given domain are distinctly recognized. This model elaborates on the method in which the world, or the domain, should be modeled to represent it comprehensively and completely enough. This model identifies the interactions that can arise within the domain. After adequate research, this researcher believes that this model of finding associations would seem to work well with knowledge systems that serve applications in social contexts such as Criminology, demographic modeling, and so on. A property of such applications is that they are more domain schema reliant rather than domain data reliant, in that, leveraging the schemas can be more useful to acquire possible result sets rather than leveraging data to acquire exact, precise results.

In the Domains of Reasoning model, the world, or the domain, is essentially considered to be a complex informal chaos. This researcher’s interpretation of Gaines’ Domains of Reasoning model segregates a given domain into distinct worlds or sub-domains as in Figure 3.6, below. [Gaines, 1987]

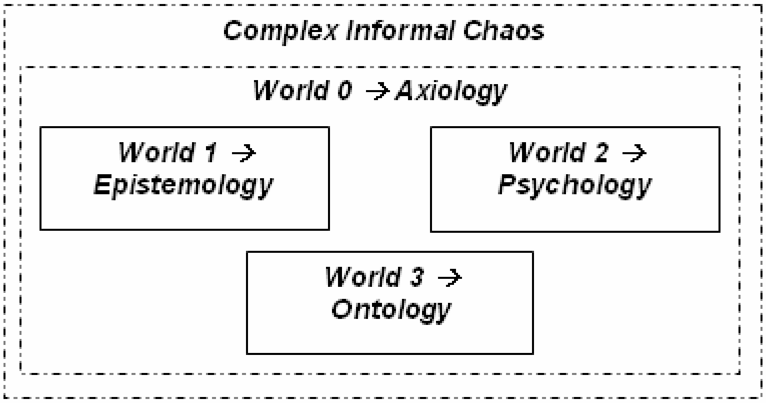


Figure 3.6.: A Compact View of the Domains of Reasoning Model

- *World 0 → Axiology:* Within the domain, distinct associations and the components involved therein are identified. Only those that are useful, based on utilitarian truth and pragmatic inference, are identified. It is called the World of Axiology since the associations identified are ascribed a truth value.
- *World 1 → Epistemology:* The representational components of a domain must convey their manifestation of physical reality, i.e. which components encompass what amount of physical reality. It is termed the World of Epistemology since the components are characterized by their value in modeling the physical domain and setting limits to the manifestation that must be represented.
- *World 2 → Psychology:* The representational components of the domain share relationships between one another. Some associations are explicit and evident, some are implicit, while others are latent or have yet to be formed. This is the World of Psychology since it is formative of the associations that are exhibited between the components of the domain.
- *World 3 → Ontology:* The components of the domain are affiliated by some common underlying characteristics, in terms of criteria that determine that they belong together. The representational components have an established nature and a system that seeks to find associations within them. This is called the World of Ontology, since the nature of the components clusters them in associative groups, thereby forming an ontology for the domain.

As an instance of the Domains of Reasoning model applied to a computational context, the Basis of Categorization model has been proposed by the researcher for a first-level categorization of the components of the corpus into possible associations and association makers. World 1, World 2 and World 3 establish these possible associations and association makers, while the associations extracted are given a truth value using the World of Axiology. The researcher's perception of the descriptions of the Worlds is thus defined as the following:

- *Axiology* – That part of the domain reliant on the nature of truth-values and truth-value judgments. [American Heritage Dictionary, 2000]
- *Epistemology* – That part of the domain reliant on the domain knowledge especially with reference to its limits and validity. [Webster's, 1998]
- *Psychology* – That part of the domain reliant on rational processes within the domain and the behavior of the domain. [American Heritage Dictionary, 2000]
- *Ontology* – That part of the domain reliant on the science of metaphysics, which investigates and explains the nature and essential properties and relations of all components, as such, or the principles and causes of their being. [Webster's, 1998]

When the components of a domain are segregated into possible associations and association makers, it becomes easier to identify and extract more objective associations between the components of the corpus.

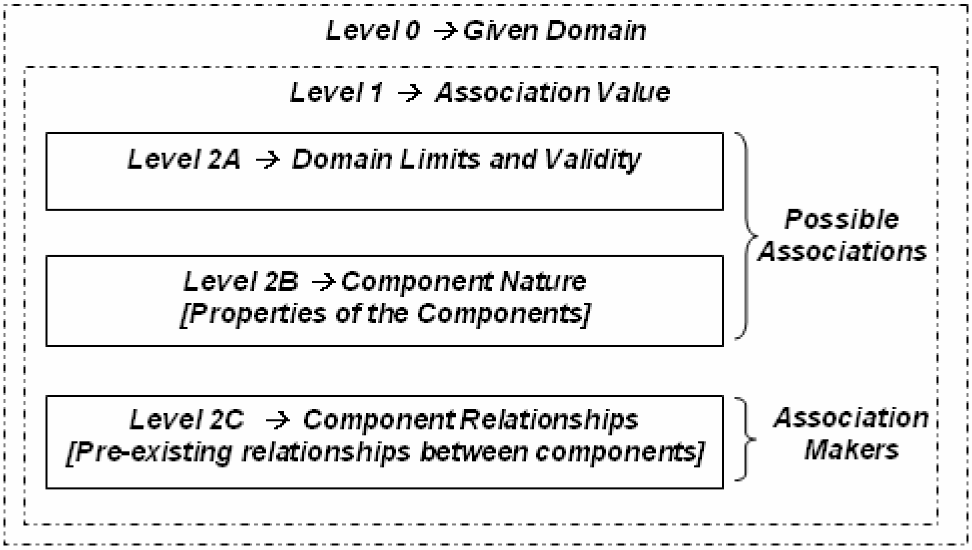

Level 0 → Given Domain
Level 1 → Association Value
Level 2A → Domain Limits and Validity
Level 2B → Component Nature
[Properties of the Components]
Level 2C → Component Relationships
[Pre-existing relationships between components]
Possible Associations
Association Makers

Figure 3.7.: The Basis of Categorization Model

The division of the domain into 'Worlds' as applied to a computational context forms the Basis of Categorization model. Figure 3.7 displays the Levels of the Basis of Categorization model. These are described as:

- *Level 0*: This indicates the domain that must be represented.
- *Level 1*: The World of Axiology, now called Association Value, stands for the goodness or the value of an association.
- *Level 2A*: The World of Epistemology, now called Domain Limits and Validity, stands for limits of the knowledge contained in the domain, i.e., the types of representational components permissible in the given domain.
- *Level 2B*: The World of Ontology, now called Component Nature, stands for the nature of being or properties of the representational components of the domain.
- *Level 2C*: The World of Psychology, now called Component Relationships, stands for the behavioral characteristics, such as causes for relationships and interactions of the components. It signifies the pre-existing relationships between the components in the corpus.

Level 1 gives a truth-value or a belief co-efficient, a weight of belief to the associations identified by Level 2. This level determines whether the resultant associations fall within the confidence intervals that have been pre-determined for the domain. Levels 2A and 2B identify components that can have possible associations, while Level 2C identifies components that are association makers in the corpus. The three parts of Level 2 – 2A, 2B, and 2C – work in conjunction to define the representational components and their associations contained in the corpus. Therefore, first-level categorization of the components in the corpus of the domain results in the following:

Given Domain
(Truth value or belief co-efficient of the association: Association Value
(Possible associations and association makers: Domain Limits and Validity + Component Nature, Component Relationships)
(Set of Criteria for Associations applied to identify relations)
)

Possible associations can exist based on the limits and validity of the domain represented in the corpus, as well as the properties and nature of the components in the corpus. Pre-existing component relationships can serve to identify and extract more associations between the components in the corpus. Now, given the above categorization of the domain – into possible associations and association makers – the Set of Criteria for Association is applied to identify and extract more objective associations between the components of the corpus.

3.3.2. The Set of Criteria for Association

The CBKR methodology can be made more expressive and be increased in analytical capacity by incorporating some measure of finding associations in the components of the corpus. This is done by, first, doing a first-level categorization of the corpus, and then, extracting more objective associations in it by applying the Set of Criteria for Categorization. In order to explain each of the eight criteria further, we shall consider:

- A Component to be a single representational element or entity,
- A Subset to be a small collection of components that are explicitly related to one another, or a group of components that share some common characteristic(s),
- A Structure to be a part of the corpus that represents just a part of the domain, but is reasonably complete in itself, and can hold its own, and
- A Partial Structure to be a part of a Structure that needs additional related partial structures to hold meaning in context.

The hierarchy of structures is that the domain is represented by a corpus of the representational schemas that model it. These schemas may be represented in different formats, and may be broken into components at their finest level of granularity. Every representation of the domain may contain a different set of components and their associations. These collated components may exist in groups called subsets. They may be categorized into structures based on relation and similarity, to objectify the corpus and cleanly categorize the components in it as much as possible. Every structure may contain partial structures or groups of components that are related to one another on the basis of some criteria. Figure 3.8 is a representation of this hierarchy.

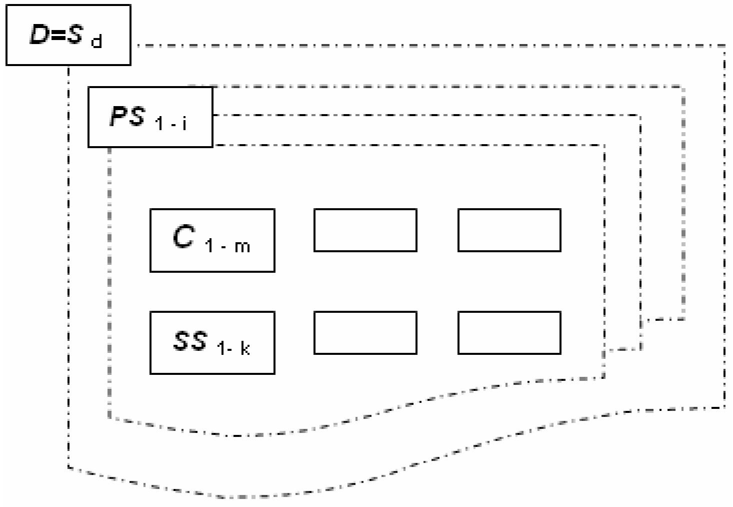


Figure 3.8.: Hierarchy of the Components in a Corpus

More formally expressed, the hierarchy can be stated as follows:

- For every domain D, there are d structures S that represent it.
- For every structure S_d, there are partial structures PS_1 to PS_i that further segregate and objectify the domain D, where i can be the degree of objectivity needed. The higher the objectivity, the higher is the refinement of corpus.
- For every partial structure PS_i, there are components C_1 to C_m that make up the partial structure. Sets of related components, sharing common attributes or behaviours can be grouped into subsets SS_1 to SS_k in order to collate components.

The criteria for association contributed by this research are descriptive and inferential in nature, descriptive because, they provide a quantitative account of the domain, and inferential because, the truth of the hypotheses suggested over the domain can be verified, and predictions made can be assumed to have a high probability of occurrence. Figure 3.9 shows the criteria that compose the Set of Criteria for Association.

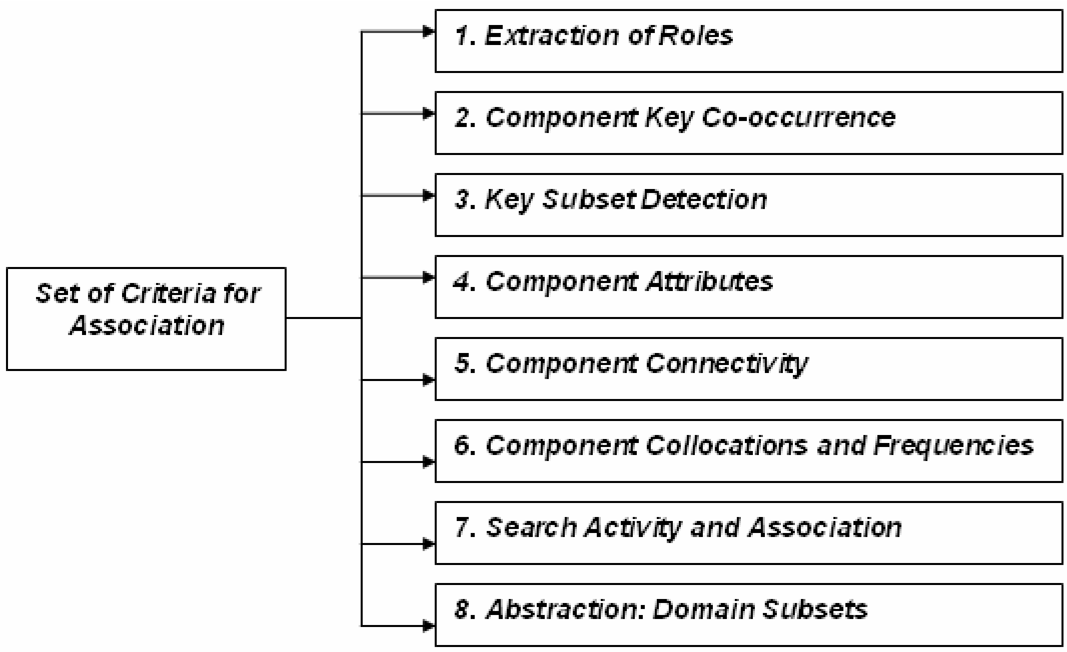

Set of Criteria for Association
1. Extraction of Roles
2. Component Key Co-occurrence
3. Key Subset Detection
4. Component Attributes
5. Component Connectivity
6. Component Collocations and Frequencies
7. Search Activity and Association
8. Abstraction: Domain Subsets

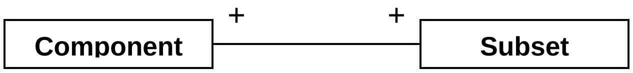
Figure 3.9.: The Set of Criteria for Association

The Set of Criteria for Association, as in Figure 3.9, is composed of the following eight criteria:

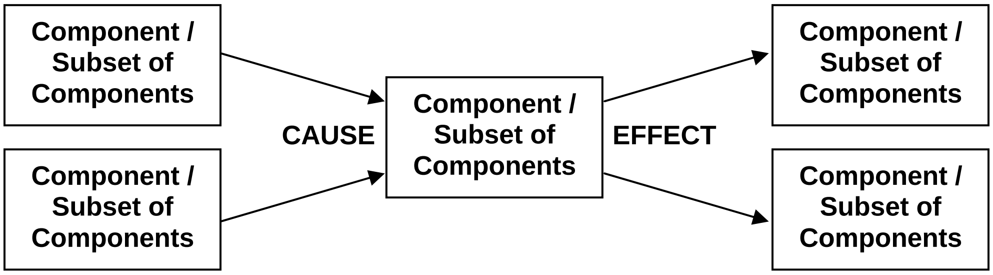
1. *Extraction of Roles*: The components and subsets of components have descriptive features that can be used to define their purpose and role within the corpus. These roles may be used in combination with one another. This criterion is sub-divided into a group that defines a given representational component. This group contains:
   - *Who*: The component name, i.e. how should the entity in question be called.
   - *What*: The component's purpose within the structure and / or the partial structure.
   - *When*: The instance of initiation and the scope of activation of the given component.
   - *Why*: The reason for initiation of the component.
   - *Where*: The place of the component in the structure and / or the partial structure, i.e., its explicit association to the structure or the partial structure.
2. *Component Key Co-occurrence*: The components in the partial structure may occur in the same representation schema, in connection with another or separately. They may be associated with one another in the same representation schema or

in different representation schemas. This criterion represents the key co-occurrences in the partial structure.

3. *Key Subset Detection*: A group of components, i.e. two or more components, form a subset. Generally, there is one representational component that is related to all others in the subset, called the key component. This criterion represents all the subsets that are associated with the key component. More than one subset can be associated with a key component, and though more than one key component can be associated with a subset, it is prudent to associate a single key component to a given subset.



4. *Component Attributes*: Generally, a component will have some characteristics in the form of some adjective qualities. This criterion represents those descriptive features that are most frequently associated with the component.
5. *Component Connectivity*: In a given representation of a domain, every component and subset of components exists because it has cause to do so – maybe, some other component or subset necessitated its existence. Also, a component or subset may affect the existence of some other components and subsets. This criterion represents the dependence between components and subsets – which components and subsets necessitate the other components and subsets to exist and how do these, then in turn, affect the existence of other components and subsets. More than one component or subset can be the cause for a given component or subset, and one component or subset can affect more than one component or subset.


Component / Subset of Components
Component / Subset of Components
CAUSE
Component / Subset of Components
EFFECT
Component / Subset of Components
Component / Subset of Components

6. *Component Collocations and Frequencies*: The representational components in a partial structure can be collocated, i.e., they can have a proper order that can be juxtaposed or arranged, especially those that commonly co-occur. This criterion represents the juxtapositions and arrangements that most frequently occur with respect to a given component, and the frequency with which they occur to form similar subsets.
7. *Search Activity and Association*: This criterion represents the associations that arise between the components and subsets when a search is conducted for a given query. The information that can be gleaned from the search activity is the association between components when a search is conducted, consecutively one after the other or in some definable pattern. This statistic trails the search activity and consequent association between components.
8. *Abstraction: Domain Subsets*: This is the one criterion in this set that quantifies the association between partial structures. It represents the abstractions of the corpus, i.e. the structure to establish domain points, which are the partial structures. Basically, depending on some criteria of association, it represents which components are related to one another and forms partial structures of those components that adhere to that association. It is very important to define the criteria of association with maximal objectivity so that the segregation of the structure into partial structures causes minimal contention among the components when they have to be classified. However, it is possible that one component may be classified under two partial structures. Through this statistic, the library of domain representations and the components of those representations are separated into groups that further serve to organize the corpus and fine tune it.

Corpus based approaches find their roots in Natural Language Processing [NLP] where, often, a corpus of text is used to elucidate linguistic theories. This can refer to anything from fairly simple string-manipulation tasks like building concordances of natural language texts, to higher-level AI-like tasks like processing user queries in natural language. Concepts such as word and phrase translation, their semantic relevance and theories of argument are generally validated using corpus of texts. Corpus based approaches also find their roots in Natural Language Interfaces [NLI], systems that allows users to access information for a given system using natural languages like English. [Androutsopoulos et al., 1995] The statistics of the CBKR methodology are founded on the paradigms of Learning Tasks, NLP, NLI and Information Retrieval [IR] [Halevy & Madhavan, 2003]. Figure 3.10 shows that the eight criteria in the Set of Criteria for Association are also based on the paradigms of Natural Language Processing [NLP] and Information Retrieval [IR] as well. They are inspired by the research presented by [Grzymala-Busse, 1986, Lau & Horvitz, 1999, Lieberman & Liu, 2002, Peat & Willet, 1991, Raghavan & Tsaparas, 2002, Rieger, 1997].

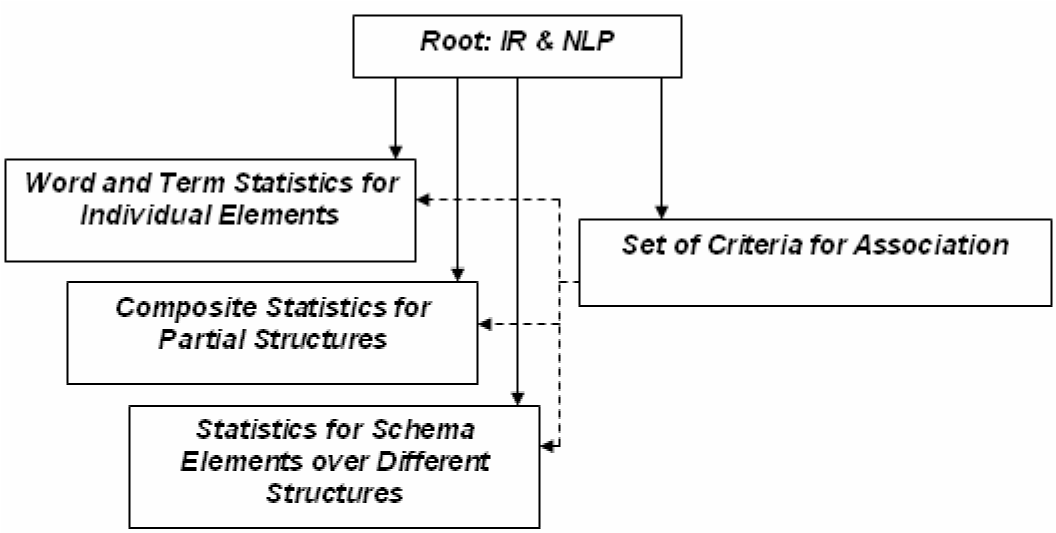

Root: IR & NLP
Word and Term Statistics for Individual Elements
Set of Criteria for Association
Composite Statistics for Partial Structures
Statistics for Schema Elements over Different Structures

Figure 3.10.: Common Roots for the Set of Criteria for Association

The diagram shows that the Statistics from CBKR, i.e., Word and Term Statistics for Individual Elements, Composite Statistics for Partial Structures and Statistics for Schema Elements over Different Structures are based on the paradigms of Natural Language Processing and Information Retrieval [Halevy & Madhavan, 2003]. Similarly, the Set of Criteria for Association is also rooted in the axioms of Natural Language Processing and Information Retrieval. The set is applied to individual representational components, partial structures composed of those components, and independent structures of the components that are the sub-sections of the main corpus. The associations acquired from the Set of Criteria for Association aid in establishing more interrelations between the components in the corpus, individually or in structure form, for use in qualitative analysis of the corpus.

3.3.3. The Relation between the Model and the Set

This research is done on the presumption that the Set of Criteria for Association should not be a random collection of criteria if the set has to be effectively applied to a corpus containing the representational components of an entire domain. These criteria must have some unifying factor so that they can work together, in conjunction, to facilitate the identification of associations between the components of the corpus. This unifying factor is the Basis of Categorization model. The three parts of Level 2 work in conjunction to define the nature, origin and limits of the representational components of the domain and their relationships. Figure 3.11 shows the relation of each criteria from the Set of Criteria for Association to Level 2 of the Basis of Categorization model. Though more than one part of Level 2 – 2A, 2B, 2C – can influence a given criterion, the figure shows only the primary influence that a part has on every criterion.

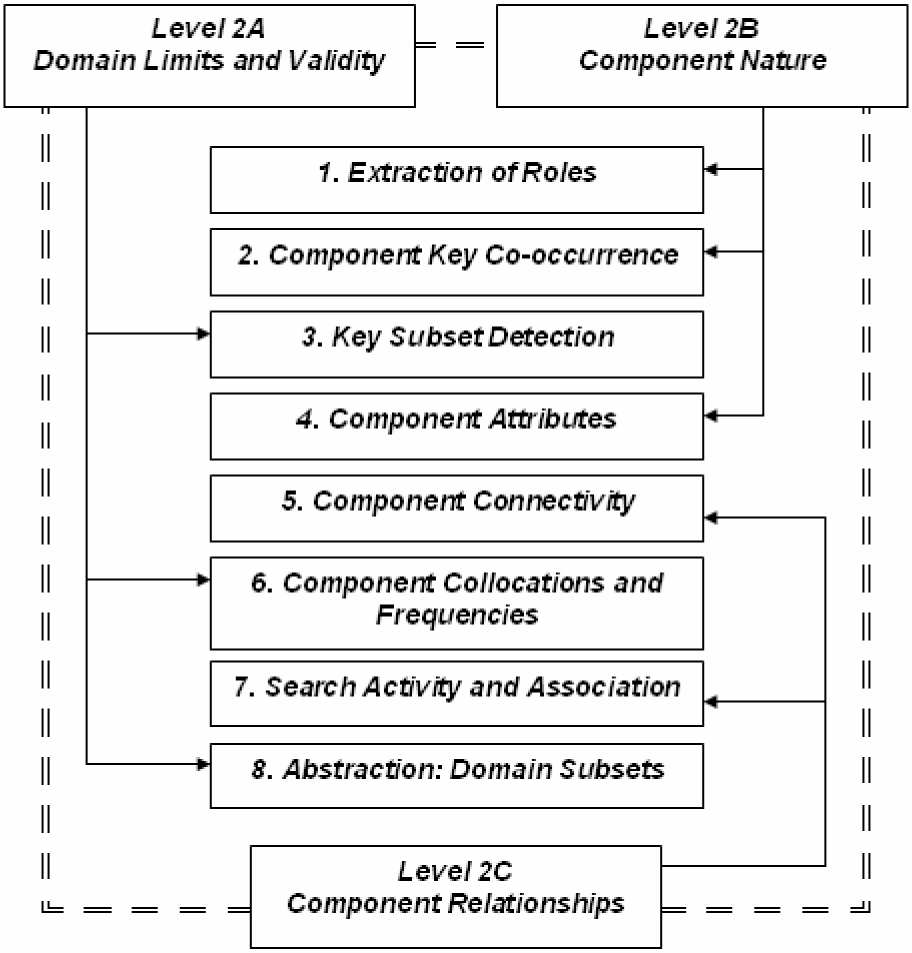

Level 2A
Domain Limits and Validity
Level 2B
Component Nature
1. Extraction of Roles
2. Component Key Co-occurrence
3. Key Subset Detection
4. Component Attributes
5. Component Connectivity
6. Component Collocations and Frequencies
7. Search Activity and Association
8. Abstraction: Domain Subsets
Level 2C
Component Relationships

Figure 3.11.: Basis of Categorization Model :: Set of Criteria for Association

Level 2A, Domain Limits and Validity, stands for nature, origin and limits of the components contained in the corpus, while Level 2B stands for the nature of being or properties of the representational components of the domain, and Level 2C stands for the behavioral characteristics, such as relationships and interactions of the components. The Set of Criteria for Association is influenced by these three parts so as to form a cohesive structure for extracting associations between all the components in the corpus. Level 2A influences the creation of the Key Subset Detection, Component Collocations and Frequencies and Abstraction: Domain Subsets criteria, as these criteria will primarily be applied to extract associations between components that model the limits and validity of the domain. Level 2B influences the need for the Extraction of Roles, Component Key Co-occurrence and Component Attributes criteria, as these criteria extract associations based on the properties of the components. Level 2C provides reason for the existence of Component Connectivity and Search Activity and Association criteria in the Set of Criteria for Association, as these criteria are primarily used to identify associations based on pre-existing relationships between components.

3.4. Employing the CBKR+ Framework

The previous sections have discussed the need for a mechanism to identify and extract associations in the CBKR methodology. An explanation of such a mechanism and its incorporation in the methodology has been provided as well. The incorporation of the Basis of Categorization model and the Set of Criteria for Association in CBKR methodology gives rise to the CBKR+ framework. As Figure 3.12 indicates, the Basis of Categorization model plus the Set of Criteria for Association, both incorporated with the CBKR Methodology, forms the CBKR+ framework.

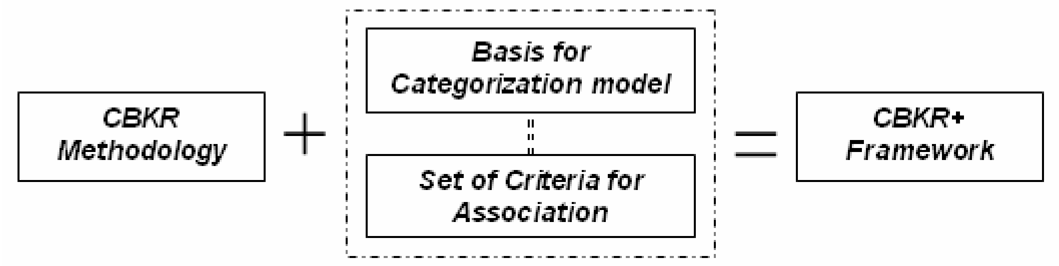

CBKR Methodology
+
Basis for Categorization model
Set of Criteria for Association
=
CBKR+ Framework

Figure 3.12.: The CBKR+ Framework

CBKR is called a methodology because it is a set of rules and procedures applied to a corpus of schemas that represent a domain. CBKR+ is called a framework because it incorporates the Basis of Categorization model and the Set of Criteria for Association with CBKR in order to enhance the technique. CBKR+ is a basic conceptional structure for classifying and organizing the components in the corpus, and is thus termed as a framework. Figure 3.13 shows the working of the CBKR+ Framework. The CBKR methodology and the Domains of Reasoning model have been covered in sections 2.3 and 2.4 respectively, while the rest forms the CBKR+ framework.

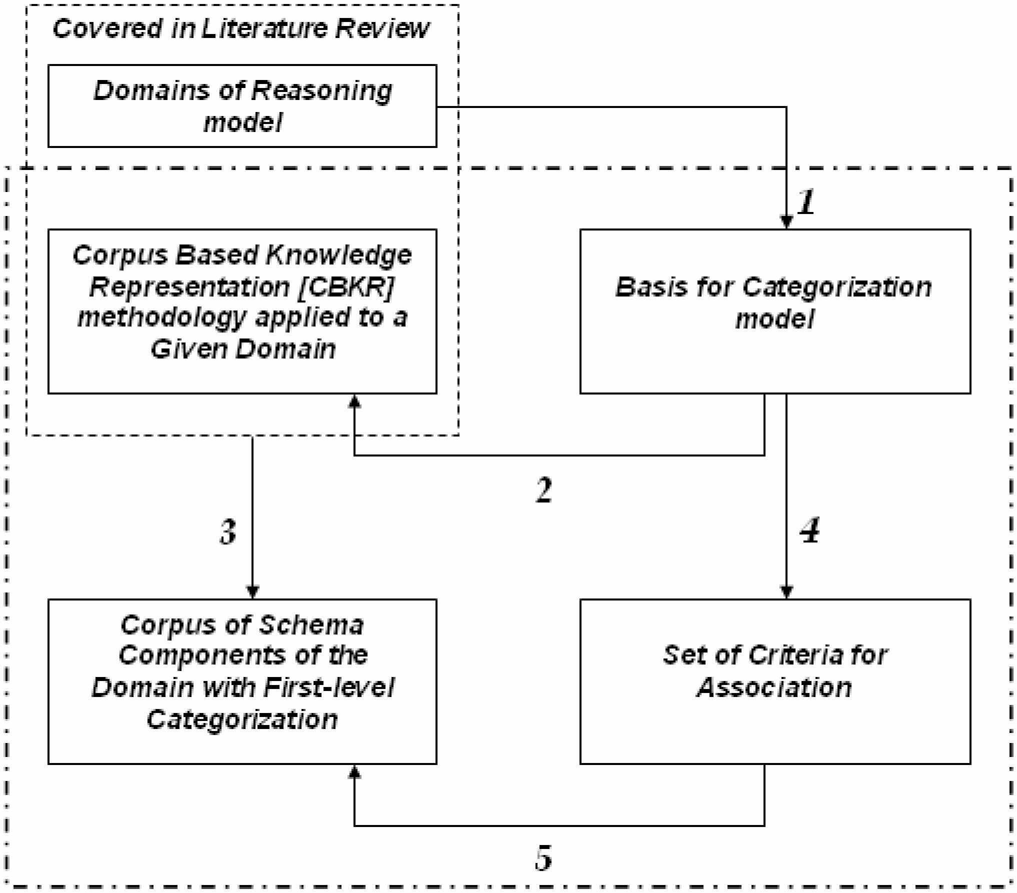

Covered in Literature Review
Domains of Reasoning model
1
Corpus Based Knowledge Representation [CBKR] methodology applied to a Given Domain
Basis for Categorization model
2
3
4
Corpus of Schema Components of the Domain with First-level Categorization
Set of Criteria for Association
5

Figure 3.13.: Working of the CBKR+ Framework

As part of this framework, Brian Gaines' Domains of Reasoning model, in the first step, instantiates the Basis for Categorization model. In the second step, the Basis for Categorization model is incorporated with CBKR. This, then, facilitates first-level categorization of schema components in the corpus into possible associations and association makers. Fourth, the Basis for Categorization model influences the creation of the Set of Criteria for Association, so as to extract associations between all possible associations and association makers in the corpus. Fifth, the Set of Criteria for Association is applied to the Corpus of Schema Components of the Domain that has undergone First-level Categorization by the Basis for Categorization model. This step identifies and extracts more objective associations in the corpus, making the CBKR methodology more expressive, and giving it greater analytical capacity. These steps form the CBKR+ framework that overcomes the drawbacks of CBKR as identified in section 3.1. The Basis of Categorization model and the Set of Criteria for

Association are 'incorporated with' the CBKR methodology. Incorporation of the Basis of Categorization model and the Set of Criteria for Association with the CBKR methodology indicates that these extensions can be used in conjunction with the original CBKR approach [Halevy & Madhavan, 2003]. The CBKR+ framework is not suggested as a replacement to the CBKR methodology, but as an enhancement to it.

Figure 3.14 shows the transition of the CBKR methodology to the CBKR+ framework proposed in this research.

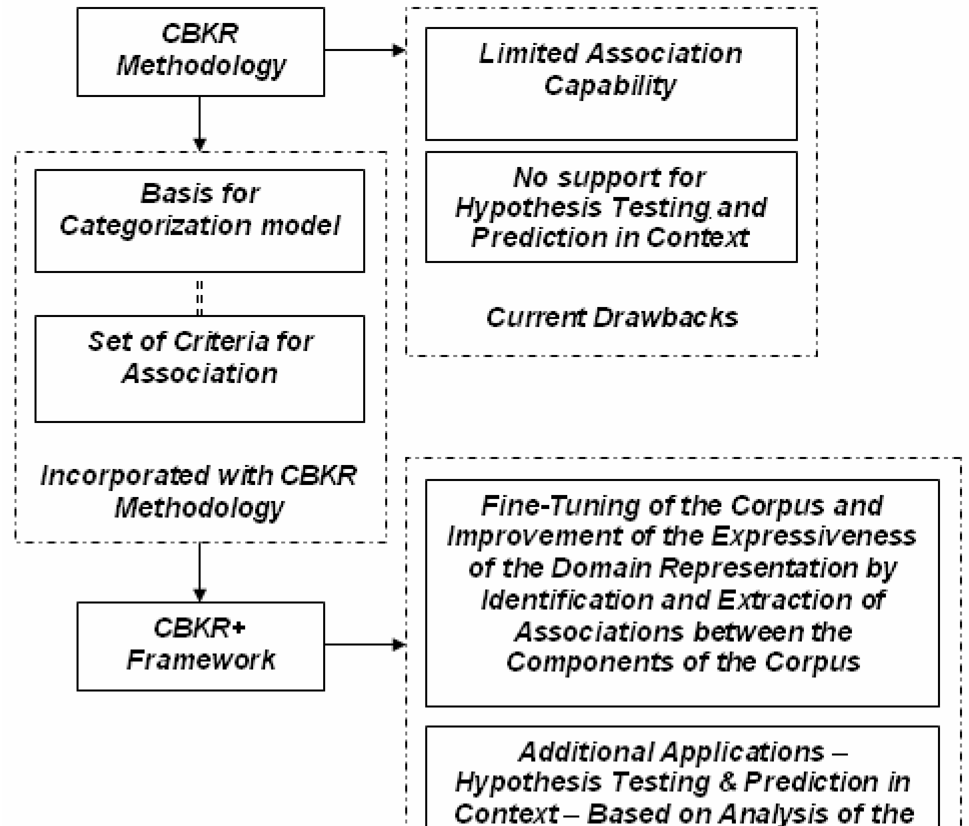

CBKR Methodology
Limited Association Capability
No support for Hypothesis Testing and Prediction in Context
Current Drawbacks
Basis for Categorization model
Set of Criteria for Association
Incorporated with CBKR Methodology
CBKR+ Framework
Fine-Tuning of the Corpus and Improvement of the Expressiveness of the Domain Representation by Identification and Extraction of Associations between the Components of the Corpus
Additional Applications – Hypothesis Testing & Prediction in Context – Based on Analysis of the Corpus are addressed
Drawbacks Addressed

Figure 3.14.: Transition of the CBKR Methodology to the CBKR+ Framework

The CBKR methodology and its drawbacks have been explained in the preceding sections. These include limited association capability and no support for hypothesis testing and prediction in context.

The incorporation of the Basis for Categorization model and the Set of Criteria for Association with the CBKR methodology forms the CBKR+ framework. This framework overcomes the drawbacks of CBKR, in that:

- *Enhanced Association Capability*: CBKR+ fine-tunes the corpus and improves the expressiveness of the domain representation using mechanisms for the identification and extraction of associations between the components of the corpus, and
- *Support for Hypothesis Testing and Prediction in Context*: CBKR+ enhances the CBKR methodology so that it can be used for applications based on analysis of the corpus, such as hypothesis testing and prediction in context.

While the framework has been designed to be as simple as possible, three of the important pre-requisites for the best possible functioning of CBKR+ are:

- The meta-strategy of identifying associations and patterns is that the training data used for this initial learning is taken from strict schemas whose components and their associations have been validated by various contexts. This forms the foundation on which additional representational components from new schemas are added to the corpus,
- The components, subsets, partial structures and structures of the domain must be as correctly recognized and carefully crafted as possible, and
- The truth values, goodness values, belief coefficients and confidence intervals must be as accurately predetermined and well defined as possible.

The following is a guideline to assist the user in employing this framework easily, once the environment to create the corpus and a set of truth values or belief coefficients that can be used for determining the confidence interval for a given criterion have been established:

- Identify the initial components, subsets, partial structures, and structures in the corpus, and establish the granularity best suited to the domain.
- Apply the Basis of Categorization model to the corpus to facilitate a first-level categorization of the corpus into possible associations and association makers.
- Then, apply the Set of Criteria for Association to identify and extract more objective associations between the components of the corpus.
- Identify a hypothesis that has to be tested, and based on the associations formed; check if the hypothesis has been tested as true or false. Perform prediction in

context using the predefined probabilities for associations and their impact on the state of the system.

One aspect to be cautious of is that this model may contain new, hidden variables that have not been accounted for earlier. As these are discovered, in the next iteration, integrate them within the structure, as part of existing partial structures or a new one altogether. Hypothesis Testing and Prediction in Context is based on observing transformations on the corpus, and identifying patterns based on the regularity of occurrence of components and associations between them.

This structure of employment has been used for the case studies in the next chapter to demonstrate the CBKR+ framework. The simplest form of diagnostic and probability-based queries is where the number of a certain type of instances is compared to the total number of instances that have occurred in the domain. If the truth-value generated by this comparison falls within the confidence interval determined for the possibility of the situation being valid, the outcome of a fashioned hypothesis can be ascertained. [Stark, 2005] As regards this research, predictability within the domain is based on the regularity of occurrence of phenomena and outcomes. The greater the regularity of these occurrences, based on the truth-values of the probabilities generated, the greater is the possibility of the correctness of the prediction. However, the one caution to using this approach is that even if the patterns recognized in the data are prominent, they may be disregarded if the amount of data is insufficient to merit confidence in the pattern. The next chapter demonstrates case studies that apply the CBKR+ framework to the domain of Criminology.

3.5. Discussion of the CBKR+ Framework

The process of fitting good models to data generated by the environment of a large domain is often computationally problematic. Efforts to alleviate this problem are the basis for developing a corpus based approach to knowledge representation. The advantages are two-fold [Halevy & Madhavan, 2003]:

- The environment of the representational components of the domain forms a corpus, in a manner similar to that of building a library. This makes the representation of the domain more complete. Associations can be identified between the representational components. Naturally occurring similar components can form groups that turn into subsets within the domain. These subsets are not isolated and can interact with other components on the basis of further associations that can be established as well. Due to the large amount of knowledge in any large domain, and the many schemas that can represent it, it becomes easier to organize domain data using a corpus of representational components that model different perspectives of the domain in many ways.

- Component collation for the corpus can be done from heterogeneous sources where the rigidity of the template for adding new schemas to the corpus does not have to be dogmatically adhered to. People who design a strictly homogeneous domain are usually the only ones qualified to make updates and additions to it. This limits the sources that can add information about the representation of the domain to the initial schema. With a corpus based approach, sources other than the original designers of that domain can update the corpus.

The characteristics of the domain define the success of the model in adapting to updates and changes in domain, and stabilizing with regard to it. The types of domains that lend themselves to the current CBKR methodology are modeled by schemas having characteristics that can be expressed in objective terms more easily, allowing the use of criteria to establish associations and patterns between the components of the domain representation. Even if the environment is too small to be reliable, the corpus based approach works, as the clustering of representational components of the domain breaks the domain down into a more organized environment. Some parts of the whole environment, now, may be more completely represented depending on the components contributed by the various schemas, and the associations formed between them. The CBKR methodology can be applied to a partial environment, thereby giving at least partial results rather than none at all. The probability of these results being accurate will obviously be lower.

The CBKR+ framework proposed in this research can be put to good use for hypothesis testing, a form of diagnosis, where a hypothesis in relation to the domain is proposed and then its truth-value is judged. Cognitive interpretation of the associations can recognize patterns in the environment based on frequency and similarity of occurrence. The CBKR+ framework can also be employed as a simple tool for prediction in context – i.e., if the associations between the components are analyzed on the basis of the observed regularities and obvious patterns in the system, the knowledge thus obtained can be leveraged to obtain constructs of what the environment might look like if certain components and their associations are changed. The hypotheses themselves maybe contemplative, and if shown to be true, can stand as predictions about the domain. This predictability is based on emergent behavior that results from observing a sufficient number of regularities in the associations of the components in the corpus of the domain. Predetermined metrics can create a spectrum of trust values that are acceptable for the given proposition within a certain confidence interval. Surprising or unexpected interaction is not considered [Gaines, 1987] in the CBKR+ framework. The outcome of the prediction is one that is expected to have a high probability of happening – the process conclusively determines whether the prediction or the hypothesis may be right or wrong. Testing hypotheses and predicting possibilities that rely on unexpected representational components and interactions require a more fuzzy form of logic.

Once proposed and explained, every model must be discussed using some performance measures. A set of criteria has been suggested by Torsun, explained in Chapter 1, for the discussion of different knowledge representation schemes. Though CBKR and CBKR+ do not belong to the traditional knowledge representation methods category, Torsun's criteria can be applied to these models as well. The following is a discussion of the CBKR+ framework according to Torsun's criteria [Torsun, 1995 as in Pesonen, 2002]:

- *Semantics*: Representations are usually intended as a medium for conveying meanings about some world or environment, and must therefore have a semantic theory that provides an account in which a particular representation corresponds to the external world or environment. The proposed framework forms organized associations between the representational components in the corpus. As a result, CBKR+ is able to provide a more coherent semantic theory of the domain, as compared to the CBKR methodology.
- *Expressive adequacy*: This criterion evaluates what part of the domain can and cannot be represented and how well. It also evaluates what part of the domain has been used and for what. The CBKR corpus of representations can address the needs of a limited class of applications. In comparison to the CBKR methodology, the CBKR+ framework applies itself to additional applications –

those that based on the analysis of the corpus, such as hypothesis testing and prediction in context.

- *Naturalness*: In some sense, all knowledge can be represented in binary form, but such a representation of the domain will be very difficult to understand and manipulate. The CBKR methodology models the domain using representational components of the representation schemas, and uses a library-like structure that can be comprehended easily. In the same spirit, the CBKR+ framework represents the domain in the form of an easily understood and implemented library structure – one that is more cohesive, better tuned and more widely applicable than CBKR.
- *Reasoning*: This is the performance metric to measure how complete and efficient is the inference mechanism and the inductive / deductive process. The proposed CBKR+ framework makes the domain representation more expressive, and endows the CBKR methodology with greater analytical capacity.
- *Primitives*: This measure evaluates what are the primitives (if any) in the knowledge representation. The CBKR methodology has an efficient but limited set of statistics to aid in the representation. The CBKR+ framework, on the other hand, supports eight criteria for identifying associations in the corpus.
- *Incompleteness*: This criterion measures the completeness of the representation of the domain. No matter what the technique, a given representation of the domain is always brittle at the edges of a model due to the vastness and heterogeneity of any domain. The CBKR+ framework identifies more associations in the corpus, thereby achieving greater completeness of the representation. Due to the dynamic nature of the world, the fine-tuning of the corpus will be an ongoing process. The CBKR+ framework represents a more complete picture of the domain than the CBKR methodology can.
- *Revisable reasoning*: Much of what one knows about the world is almost always true. This metric evaluates the methods that formalize and compute this truth. The proposed CBKR+ framework does not support a fuzzy form of logic and reasoning, but identifies associations between the components in the corpus. It supports confidence intervals and degrees of belief that reflect the real world are used as part of the analytical process.

- *Flexibility:* Any given domain is dynamic and susceptible to change. It is essential that the representation system have the ability to add new representations of the domain and update the corpus of representational components with ease. The CBKR+ framework has the ability to accept more schemas in the corpus, and categorize the new components according to the patterns of association observed in the corpus, thereby making the system more flexible than the CBKR methodology.

3.6. Chapter Summary

In this chapter, the reasons for the need of the CBKR methodology to support some mechanism for forming associations among the representational components were explained. Towards this goal, the methodology was enhanced by incorporating the Basis of Categorization model and the Set of Criteria for Association in it. This formed the CBKR+ framework. The influence of the Basis of Categorization model on the Set of Criteria for Association was discussed. These concepts were collated to explain the working of the CBKR+ framework. Finally, the CBKR+ framework was discussed on the basis of a set of performance criteria suggested by Torsun.

Chapter 4 – Application of the CBKR+ Framework

The possession of knowledge does not kill the sense of wonder and mystery.
There is always more mystery.
~ Anais Nin (1903 - 1977), Writer

4.0. Chapter Introduction

This chapter presents a demonstration of the CBKR+ framework through case studies. The seminal example from CBKR is shown as a preliminary case study for CBKR+. The framework is also applied to the domain of Criminology as a physical world case study.

This chapter is organized as follows: general comments about the application of the proposed framework are followed by an illustration of the examples of the CBKR methodology, and an indication of how the CBKR+ framework can enhance the capability of the methodology. The working of the CBKR+ framework and the use of the framework to support the applications of Hypothesis Testing and Prediction in Context is then demonstrated by applying it to a physical world domain of Criminology.

4.1. Application of the CBKR+ Framework

The CBKR+ framework proposed in this research collates a library of representational models or schemas of a given domain. These models are heterogeneous as they represent different perspectives and different aspects of the domain using various schemas. CBKR and CBKR+ are not proposed as replacements for traditional knowledge representation, but are suggested as complementary methodologies that use schemas of domain representation, rather than data elements and fragments of specific knowledge. This collation of schemas and the representational components contained therein in a library-like structure leads to the formation of a corpus. The representation of the domain is macroscopic in character. This is because the domain is represented in its structural form through schemas instead of the content of the domain, such as instances of data elements and fragments of specific knowledge. Specific data instances within each component form the microscopic view of the domain, which is dealt with using rigorous and more complex NLP techniques beyond the scope of this research. The CBKR methodology has limited capability to form associations between the components of the schemas in the corpus. The CBKR+ framework identifies and extracts more associations in the corpus. The associations identified serve to build a clearer, more comprehensive representation of the domain. The CBKR+ framework enhances the organization of the schema components in the corpus, thereby fine-tuning the corpus, and increasing the range of applications that can use this form of representation.

Through the framework, the Basis of Categorization model creates a first-level, elementary categorization of the given domain corpus into possible associations and association makers. Then the Set of Criteria for Association is applied to this corpus to identify and extract more objective associations between the components of the corpus. Finding associations in the corpus with the CBKR+ framework is conducted on the presumption that there are patterns in the corpus that can be identified. These patterns are identified on the regularity of occurrence of the components in the corpus and the relations that exist between them. Representational models of demographic and ethnographic domains such as environmental monitoring, marketing strategies, and criminology have a higher possibility of containing potential patterns. [Halevy & Madhavan, 2003] This is because there is the possibility of (1) a large number of schema representations created by a variety of experts due to the large size of such domains, (2) similar data content that is modeled differently through many schemas giving rise to the formation of a corpus, (3) the use of such domains to acquire approximate or partially correct result sets rather than strictly accurate ones, and (4) the use of such domains for multiple applications that are based on similar representations of the domain. Applications that leverage such domains must also be more domain schema reliant rather than domain data reliant for the effective use of the CBKR+ framework. The result sets of such applications are gradually rather than rigidly correct, where approximate answers can be more useful than classically correct ones. The mechanism of identifying associations by the CBKR+ framework first works inductively and then deductively. It is inductive as specifics about the representational components of the corpus are used to arrive to general conclusions about all of them through analyzes of the corpus. It is deductive because general warrants are used to test the veracity of specific hypotheses and claims made about the domain, where results are used to judge the truth-value of these claims using probabilities and predetermined confidence levels.

Figure 4.1 is a graphical representation of the application of the CBKR+ framework. Various representational or structural schemas of a domain are brought together in a corpus. These schemas are composed of components, where a component is the finest granularity of a schema, and is determined by the designer of the schema. The Basis of Categorization model and the Set of Criteria for Association are applied to this corpus. This is the application of the CBKR+ framework.

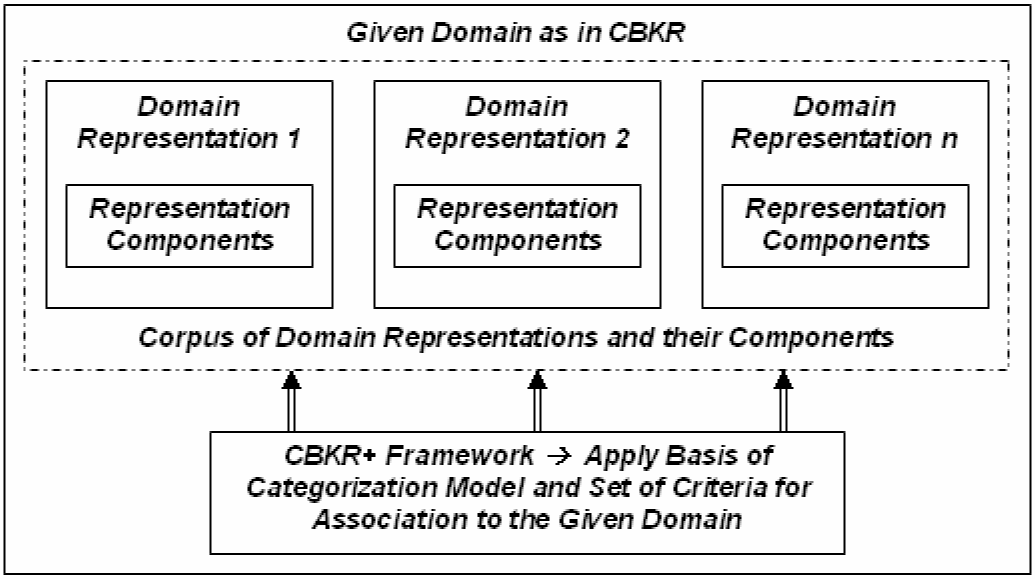

Given Domain as in CBKR
Domain Representation 1
Representation Components
Domain Representation 2
Representation Components
Domain Representation n
Representation Components
Corpus of Domain Representations and their Components
CBKR+ Framework → Apply Basis of Categorization Model and Set of Criteria for Association to the Given Domain

Figure 4.1.: High-Level Application of the CBKR+ Framework

This chapter applies the CBKR+ framework postulated in the previous chapter the example form the CBKR methodology, as well as to the domain of Criminology. The aim is to test hypotheses with regard to the domain of Criminology on a corpus of various representations of the domain, and also to predict if certain possibilities regarding the domain are true.

4.2. Explaining Hypothesis Testing and Prediction in Context

The setting up and testing of hypotheses is essentially stating the proposition of some theory that is believed to be true or is to be used as a basis for argument, but that has not been proved. The hypotheses may be statements about population parameters like expected value and variance, or statements about the form of a characteristic of interest. Outcomes of hypotheses tests are either the rejection of a hypothesis or an acceptance of it, i.e. they may either be true or false. [Easton & McColl, 2005] This research considers a hypothesis in the following perspective: Given the corpus of a domain represented by various schemas, there may be a need to know whether certain propositions and assumptions about the domain hold true. A hypothesis is framed using the components concerned and the associations determined between them. Through the analyses and diagnoses of the corpus, if the certain related associations between the components are identified, the hypothesis is said to be true, or the hypothesis is accepted. Prediction in context expects the domain to be in a certain state given the associations (or the context) that affect that state. A certain number of occurrences of particular associations are related to a certain probability of the state of the domain based on observed patterns in the corpus. If this certain number is achieved, there is a high probability of the domain being in the said state. Of course, external factors can influence the prediction and lower the possibility of the domain having the predicted structure. The future addition of any new components and their associations with other components will be guided by the categorization facilitated by the Basis of Categorization model, the associations identified by the Set of Criteria for Association, the hypothesis tested, and the predictions made in context of the domain.

As an example, say, the components of a representational schema S_1 are associated to those of another schema S_2, both of which differently model domain D, since D is large and is not modeled only by one schema. These schemas and their components form the corpus. The corpus thus has components C_i that represent the domain. If the components of S_1 are C_a, C_b, C_c, and those of S_2 are C_a, C_c, C_d, then the corpus contains collated components C_a, C_b, C_c, C_d. Now, if the components of S_3, another representation schema, C_e, C_f, C_g are added to the corpus, it will contain C_a, C_b, C_c, C_d, C_e, C_f, C_g. If the associations $C_e \equiv C_a$, and $C_f \equiv C_b$ exist, then the corpus contains C_a / C_e, C_b / C_f, C_c, C_d, C_g, which will still provide a more comprehensive representation of D than those of S_1, S_2, or S_3 alone. If S_4 is now introduced, and certain representational components and their associations correspond to similar ones in the corpus, they will be added to the corpus on the basis of these similarities identified on the basis of the regularity of their occurrence in the corpus. Hypotheses are tested on the corpus containing all the representational components thus far. The pre-determined values for probabilities, confidence intervals and truth values are selected after extensive study of the domain, and are modeled within the system and updated with regard to inclusion of new information by a domain expert. If a new hypothesis is accepted or a prediction holds true, its representational components and the associations between them become part of the corpus as well.

As concerns the physical world validation of the proposed CBKR+ framework, the domain of Criminology can be represented using various schemas and models, which can be generated using literature descriptions of the domain. [Aked, 2005, Selie, 1996, Petherick, 2005]. The preliminary associations between the representational components in the corpus can be established in the spirit of [Evermann & Wand, 2005, Solan et al., 2005]. Then, the CBKR+ framework is used towards fine-tuning the corpus of representational components of the domain in order to arrive at a more comprehensive representation of the domain. Given the domain of Criminology, for example, one schema is of text sentence type, and models the domains from the point of view of defendant characteristics, i.e. Domain Representation 1. Domain Representation 2 is a concept map-like schema that models the domain from a general arrests and convictions point of view. Together, they represent a more comprehensive view of the domain. The presence of certain offences or offender characteristics and the associations between them in the corpus, such as the nature and seriousness of the charge, the extent of prior arrests and convictions, a history of drug addiction, and so on can be used to correctly hypothesize that defendants will commit new offences on release. [Petherick, 2005] These defendant characteristics may have been part of a single domain representation, or may have been collated from different representations. Figure 4.2 shows the corpus for the domain with representational components and their associations that have been collated from two different representational schemas.

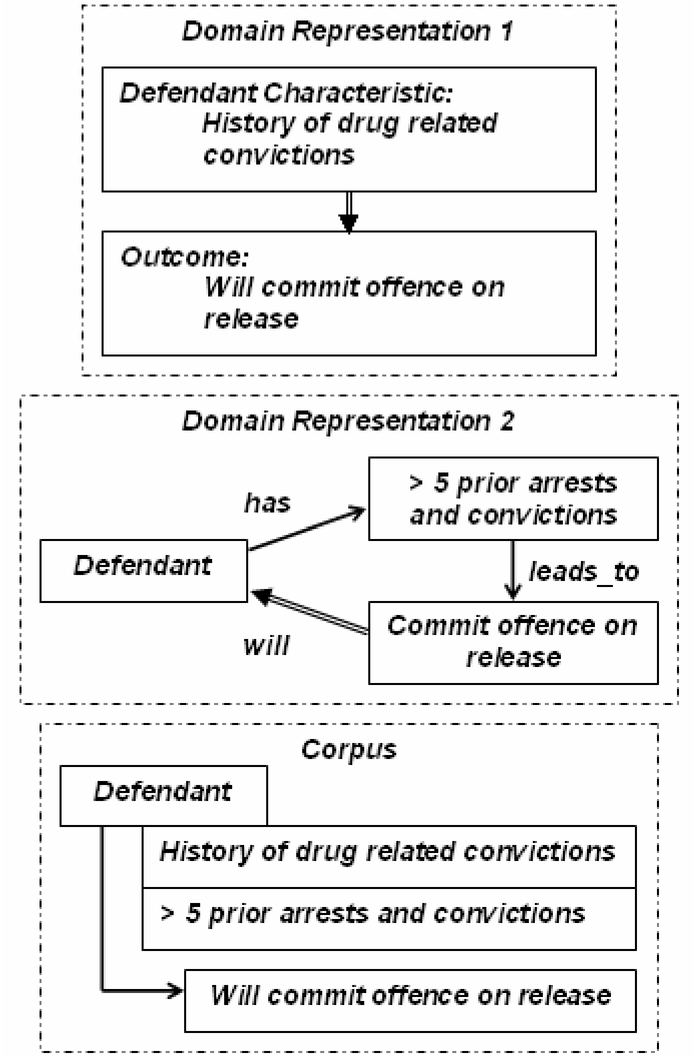

Domain Representation 1
Defendant Characteristic:
History of drug related convictions
Outcome:
Will commit offence on release
Domain Representation 2
has
> 5 prior arrests and convictions
Defendant
leads_to
Commit offence on release
will
Corpus
Defendant
History of drug related convictions
> 5 prior arrests and convictions
Will commit offence on release

Figure 4.2.: Snapshot of the Corpus with Multiple Representations

Given the characteristics of a defendant, it can now be hypothesized if a certain type of defendant will be likely to commit crimes on release, and it can be predicted which type of defendants may do so. While running these criteria on the corpus of representational components of the domain of Criminology for hypothesis testing and prediction in context, it is better to be overcautious with predictions, stating more often that a criminal type will be likely to commit offences, even if they may actually not. The penalties for failing to identify a dangerous type of individual correctly can have serious repercussions. [Petherick, 2005]

Now, introduce Domain Representation 3, which models the domain from another defendant related point of view. This schema is created using a flow chart-like structure or even pseudo code.

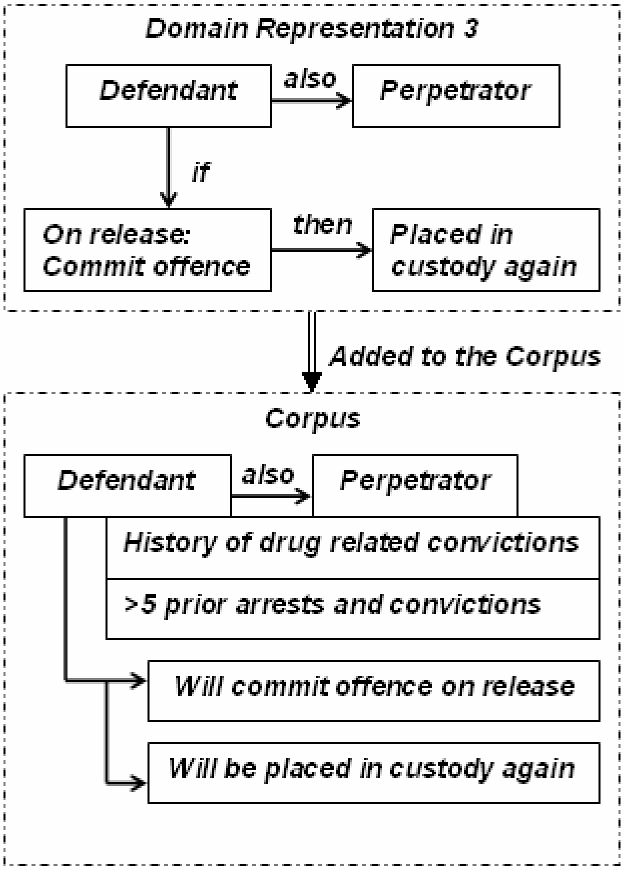

Domain Representation 3
Defendant
also
Perpetrator
if
On release: Commit offence
then
Placed in custody again
Added to the Corpus
Corpus
Defendant
also
Perpetrator
History of drug related convictions
>5 prior arrests and convictions
Will commit offence on release
Will be placed in custody again

Figure 4.3.: Additional Components and their Associations in the Corpus

Figure 4.3 now indicates that the representational corpus has additional components and associations between them added to it by the inclusion of the components of a representational schema that modeled another aspect of the domain. The component defendant is also called perpetrator. The component placed in custody indicates that the defendant will be re-arrested after committing another offence on release. A prediction in context might operate on the following assumption: 75% of offenders released from prison who have a history of drug related convictions and who were originally incarcerated for drug related offences, will go on to commit further crimes and again be placed in custody. The outcomes can be false positives, where defendants are identified as future risks when they are actually not, false negatives, where it is identified that certain individuals will not act dangerously but actually they do, true positives which are correct assessments that individuals will be dangerous, or true negatives which are correct assessments that the individuals will not be dangerous. [Petherick, 2005]

4.3. Applying the CBKR+ Framework to an Example of the CBKR Methodology

CBKR can do simple query processing for factual information and similarity queries. Results to applications may be ranked instead of determining one exact result on the basis of Boolean conditions on data. The results may also summarize sets of schema fragments. Consider the case of factual queries, where the corpus consists of schema components. For a given query, the CBKR methodology can return a set of components applicable to the query. This shows that the components form a cluster, but the methodology is incapable of specifying the exact relationship between the components. Consider a schema representation for the domain of education from the student point of view. The corpus contains the components <GPA>, <StudentID>, <Student>, <value>, and <address>. For a given factual query to find the domain representation of the GPA of a Student, CBKR may return a result set that includes the components <Student, StudentID, GPA, address> to specify the schema description. It cannot however establish the relation between the components to explain why this particular cluster was returned. The CBKR methodology is also applied to similarity queries. In a corpus, similar facts may be expressed in different ways. Similarity queries are based on leveraging these different expressions of similar facts. Consider a schema that represents the domain of vehicles from the class of luxury cars point of view point of view. This part of the corpus contains the components <Car-Review>, <Lexus>, <Toyota>, <LuxuryClass>, <VeryGood>, <goodTires>, <luxury>. For a given similarity query to find the domain representation of the review status, CBKR may return result sets that include the cluster of components <LuxuryCar, Lexus, Toyota> or <Car-Review, Lexus, LuxuryClass, VeryGood, goodTires> to specify the schema description. The CBKR methodology, however, still does not specify the associations between the components, nor is it possible to test hypotheses on the corpus or perform prediction in context.

The CBKR+ framework applied to the above examples can (1) categorize the components in the corpus into possible associations and association makers using the Basis of Categorization model, and (2) identify and extract particular associations between the components using the Set of Criteria for Association. The goal is to fine-tune the corpus and specification of these relations, and by extension, make the methodology applicable to a broader class of applications. The domain examples are very limited, and may result in overlaps of associations. Initially, the components undergo a first-level categorization to be segregated into possible associations [Levels 2A and 2B] and association makers [Level 2C]. For the examples from CBKR, the components (1) <GPA>, <Student>, <StudentID>, <value> and <address>, and (2) <Lexus>, <Toyota>, <LuxuryClass>, <VeryGood>, <goodTires>, and <luxury> fall in levels 2A and 2B; while (2) <Car-Review> falls in 2C.

Given that, by applying the Set of Criteria for Association, more relations are observed. The criterion of Extraction of Roles describes the 'who, what, when, why and where' of each component. The common values of who, what, and so on helps to create partial structures in the corpus, where all the components in each partial structure are associated due to their common value. From (1), a <Student> is 'who' <GPA>, <StudentID>, and <value> are 'what', and <address> is 'where'. From (2), <Lexus>, <Toyota>, <LuxuryClass>, <goodTires>, and <luxury> are 'what', while <Car-Review> is 'where', and <VeryGood> is 'why'. The criterion of Component Key Co-occurrence associates components that may occur in the same representation schema, and that may be associated with one another in the same or in different representation schemas. For (1), <GPA>, <StudentID>, <Student>, <value>, and <address> occur in the same schema, and for (2) <Car-Review>, <Lexus>, <Toyota>, <LuxuryClass>, <VeryGood>, <goodTires>, and <luxury> occur together. The Key Subset Detection criterion creates an association between subsets or groups of components that are most associated with an identified key component. For the examples above, (1) for the key component of <Student>, the subset is formed of the components <GPA>, <StudentID>, <value>, and <address>, and (2) for the key component <Toyota>, the subset formed includes <Car-Review>, <LuxuryClass>, <VeryGood>, and <goodTires>. The Component Attributes criterion identifies associations based on the descriptive features that are most frequently associated with a component. For (1) <StudentID> is the feature associated with <Student>, and for (2) <Lexus> and <Toyota> are frequently associated with <Car-Review>. The dependence between components and subsets relating to cause and effect form associations based on the criterion of Component Connectivity. In the representations from the CBKR methodology, (1) <StudentID> leads to <address>, and (2) <LuxuryClass> leads to <Lexus> described in <Car-Review> as <VeryGood>. Component Collocations and Frequencies identifies associations based on juxtapositions and arrangements that most frequently occur with respect to a given representational component. For (1), <GPA> is frequently seen to occur with <value>, <StudentID> and <address>, and (2) <Car-Review> occurs with <VeryGood>, <goodTires>, while <LuxuryClass> occurs with <Lexus>, <Toyota>. The associations that arise between the components and subsets when a search is conducted for a given query, consecutively or in some definable pattern are based on the criterion of Search Activity & Association. For (1), searches for <StudentID> : <Student>, and <StudentID> : <address> result in <StudentID> : <Student>, <address>, and (2) searches for <Lexus> : <LuxuryClass>, <Toyota> : <LuxuryClass>, and <LuxuryClass> : <Car-Review> result in <Car-Review> : <LuxuryClass>, <Lexus>, <Toyota>. The Abstraction: Domain Subsets criterion identifies which components form domain points through partial structures. In example (1), <GPA>, <StudentID>, <address> are formed around <Student>, and in (2), <LuxuryClass>, <VeryGood>, <goodTires> are formed around <Car-Review>.

As seen from the above, CBKR could only find components associated as <Student, StudentID, GPA, address>, and <LuxuryCar, Lexus, Toyota> or <Car-Review, Lexus, LuxuryClass, VeryGood, goodTires> to specify the schema description. On the other hand, by using the CBKR+ framework more specific associations have been identified and extracted as seen in the examples above.

4.4. Physical World Application – Criminology

Criminology is an investigative tool available for various people and processes in criminal justice, and is used to investigate a crime. Among its many interpretations, it "is an analytical process resulting in the description of personal characteristics as differentiating and behavioural units, which are, because of their relative durability and recognition within personal activity and structure, an adequate basis for the description of the most probable perpetrator of the proceeded criminal offence."

[Selie, 1996, Page 3] While not the answer to all problems, or a complete solution, this tool can be used to reduce the possible number of suspects in an investigation, and to focus the investigation on the suspects identified. Criminology also assists in linking potentially related crimes by identifying and narrowing possible crime scene indicators and behaviour patterns, and assessing the potential for increase of petty criminal behaviour to more serious or more violent crimes. [Petherick, 2005] The use of this research to the domain of Criminology is particularly focused on the aspects of identifying types of suspect pools, types of suspects and conditions where there is potential for increase of petty criminal behaviour to more serious or more violent crimes, types of behaviour patterns with regard to crimes committed, and so on. The greater goal is to generate automatic records periodically, so that justice authorities can watch over types of high-risk situations and individuals without labor-intensive collation and analysis of information acquired by a case-by-case analysis of the crime domain. Specific instances involving specific individuals and specific circumstances have to be dealt by a human task-force due to the sensitivity and volatility of the nature of the domain.

The psychological perspective of the domain of Criminology says that "the applicability of psycho-dynamically oriented theoretical concepts and the capability of the investigator to recognize concrete behavioural patterns and their interpretation in a given situation are based on the postulates of psychological connection between personal characteristics and behavioural manifestations." [Selie, 1996, Page 1] In simple terms, that means that any situation and behaviour must be interpreted keeping the personality – desires, expectations, motives and interests, and circumstantial conditions in mind. Criminology then becomes a tool for the creative synthesis of knowledge towards identifying the important personal characteristics and behavioural patterns of perpetrators. [Selie, 1996]

According to experts from the field, tools for Criminology use quantified factors from the domain in a systematic manner, such that the profiling will not point to a specific individual, but to a 'type' of offender. The quantified factors include, but are not limited to, information regarding mental health history, possible physical appearance, work history and schedule, possible residential location and marital status. Conclusions that affect specific individuals are based on the specific information gathered from the crime scenes, and an analysis of the individual's behaviour within the context of the crime. It is claimed, though, that the scientific study of crime and criminal behavior will never replace traditional police work. [Petherick, 2005] A multidisciplinary approach to criminal investigation of classical investigative procedures coupled with good profiling systems will provide a better platform on which to fight crime. In Criminology, the domain of application of this research, known offender populations are studied and broad offender groups known as typologies are defined. These are clusters of types of individuals or circumstances or any aspect relevant to the domain by virtue of shared or similar characteristics, such as the motive for offences committed by the group, the behavioural precursors to offending, and different probabilities for repeat offending. The crimes of unknown offender types are then compared to these groups to identify and confirm specificities. Police work then applies the typologies to the assessed crime scene or behaviour of the offender during the offence. Typologies generalize behaviour based on past similar offenders, are not equipped to provide any detailed offender information, and should be only used as an initial investigative tool. Typologies produce a list of characteristics likely to be possessed by unknown offenders by virtue of their similarity to the known offender groups. [Petherick, 2005]

4.4.1. Domain Description

The validation of this research is based on a collection of typologies from the domain of Criminology. It is assumed that a collection of typologies will represent the domain more completely than any one alone. Each typology is a representation of a certain aspect or perspective of the domain. The CBKR+ framework does not profile single and exact cases. It generates possibilities that serve to test hypotheses and predictions. The components for the formation of the corpus could be acquired from the schemas of domain representations formed by (1) the analysis of solved cases, (2) discussions with perpetrators, condemned and otherwise, (3) interviews with offenders in correctional facilities about their backgrounds, (4) study of crimes, crime scenes, and victims, (5) official sources of information, such as court transcripts, police reports, and psychiatric and criminal records, evidence logs, evidence submission forms and autopsy reports, (6) interviews of witnesses and neighbours, (7) victim background information, (8) literature on criminal psychology and victimology, forensic sciences, security sciences, and so on. [Petherick, 2005, Selie, 1996] The schemas in Figure 4.4 have been developed by the referenced sources on the basis of the above. The various forms in which the domain may be represented include, but are not limited to text descriptions, tabular conjunctional values, record sets, XML templates, generation charts, classification trees, concept maps, and so on.

Figure 4.4 is a high-level representation of the corpus. The corpus contains twelve schema representations of the domain that have been acquired from the referenced academic sources.

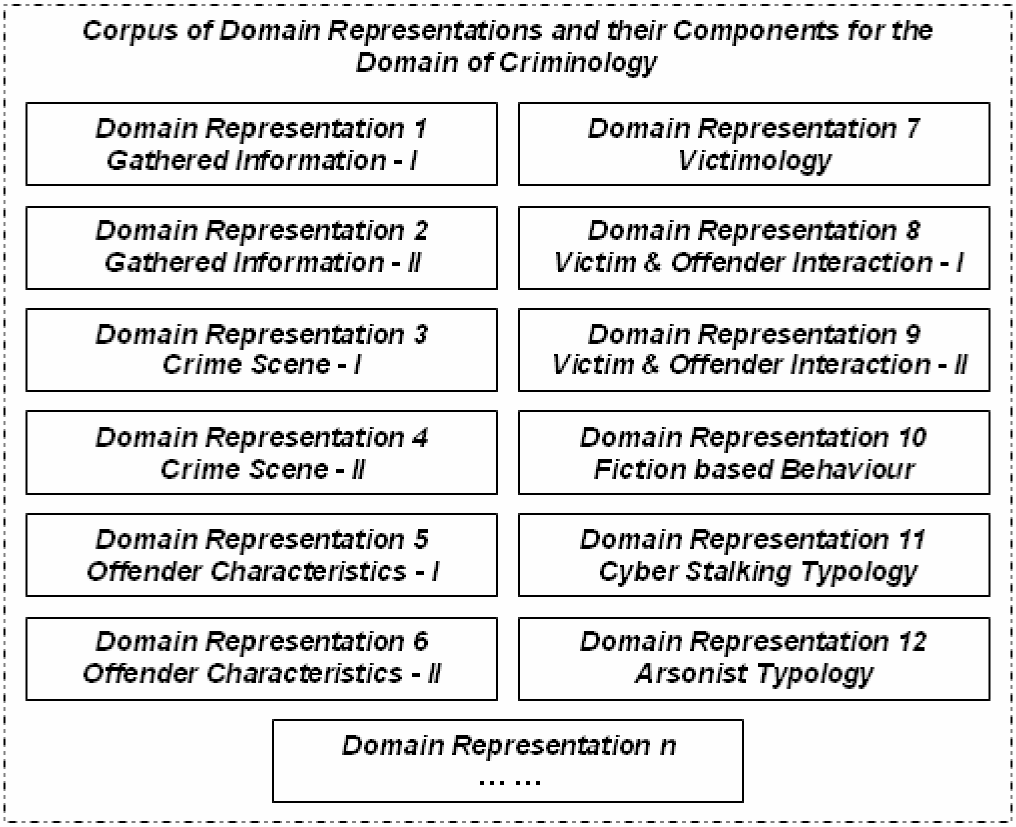

Corpus of Domain Representations and their Components for the Domain of Criminology
Domain Representation 1 Gathered Information - I
Domain Representation 7 Victimology
Domain Representation 2 Gathered Information - II
Domain Representation 8 Victim & Offender Interaction - I
Domain Representation 3 Crime Scene - I
Domain Representation 9 Victim & Offender Interaction - II
Domain Representation 4 Crime Scene - II
Domain Representation 10 Fiction based Behaviour
Domain Representation 5 Offender Characteristics - I
Domain Representation 11 Cyber Stalking Typology
Domain Representation 6 Offender Characteristics - II
Domain Representation 12 Arsonist Typology
Domain Representation n
… …

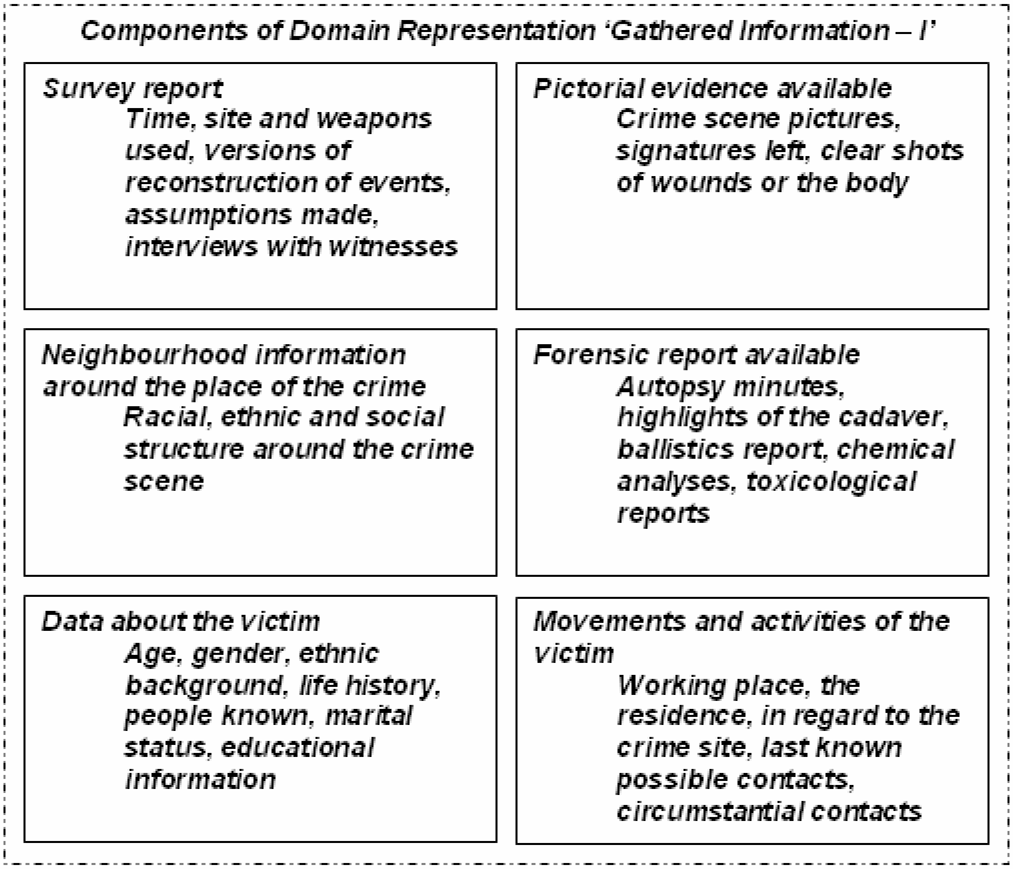
Figure 4.4.: Corpus of Collated Representational Components for the Domain

Each representation is explained in the following sections in terms of its components. The many aspects and perspectives of the domain include, but are not limited to, the typologies described in the following sections. Overall, domain representations 1 and 2 model the same aspect of the domain in different perspectives, as do representations 3 and 4, 5 and 6, and 8 and 9. Representations 7, 10, 11 and 12 model different aspects of the domain individually. There exist overlaps in the similarity of the aspects modeled by representations 9 and 11. This figure also indicates that there is scope for addition of other representations as well.

4.4.1.1. Representation 1: Domain modeled on 'Information Gathered – I'

In [Selie, 1996], a set of components of the representation of the domain using information gathered has been mentioned. This set forms a representation of the domain from one given perspective. The representational components are:

- Survey report: Data on time, site and weapons used, versions of reconstruction of events, assumptions made, interviews with witnesses.
- Pictorial evidence available: Crime scene pictures, signatures left, clear shots of wounds or the body.
- Neighborhood information around the place of the crime: Data on the racial, ethnic and social structure around the crime scene.
- Forensic report available: Autopsy minutes – intuitive observations of dissector, highlights of the cadaver – contusions, bruises, wounds, ballistics report, chemical analyzes, toxicological reports.
- Data about the victim: Age, gender, ethnic background, life history, people known, marital status, educational information.
- Movements and activities of the victim: Data acquired at the working place, the residence, in regard to the crime site, last known possible contacts, circumstantial contacts.



Components of Domain Representation 'Gathered Information – I'

Survey report
Time, site and weapons used, versions of reconstruction of events, assumptions made, interviews with witnesses

Pictorial evidence available
Crime scene pictures, signatures left, clear shots of wounds or the body

Neighbourhood information around the place of the crime
Racial, ethnic and social structure around the crime scene

Forensic report available
Autopsy minutes, highlights of the cadaver, ballistics report, chemical analyses, toxicological reports

Data about the victim
Age, gender, ethnic background, life history, people known, marital status, educational information

Movements and activities of the victim
Working place, the residence, in regard to the crime site, last known possible contacts, circumstantial contacts

Figure 4.5.: Domain Representation based on Gathered Information - I

Components such as survey report, pictorial report, forensic evidence available, etc. may be considered to be partial structures or individual representational components in the corpus, depending on the granularity desired.

4.4.1.2. Representation 2: Domain modeled on 'Information Gathered – II'

In another model to represent the same perspective of the domain, as above, the following components are used: age, race, ethnicity, marital status, vocational stability, education, socioeconomic status, developmental family, juvenile arrest history, adult arrest history, mental health status, intelligence and psychological scores on standardized tests, level of arousal, history of violence, and a signature if any. [Selie, 1996] A signature is a set of behaviours that fulfill a psychological or physical need in the offender. [Petherick, 2005]

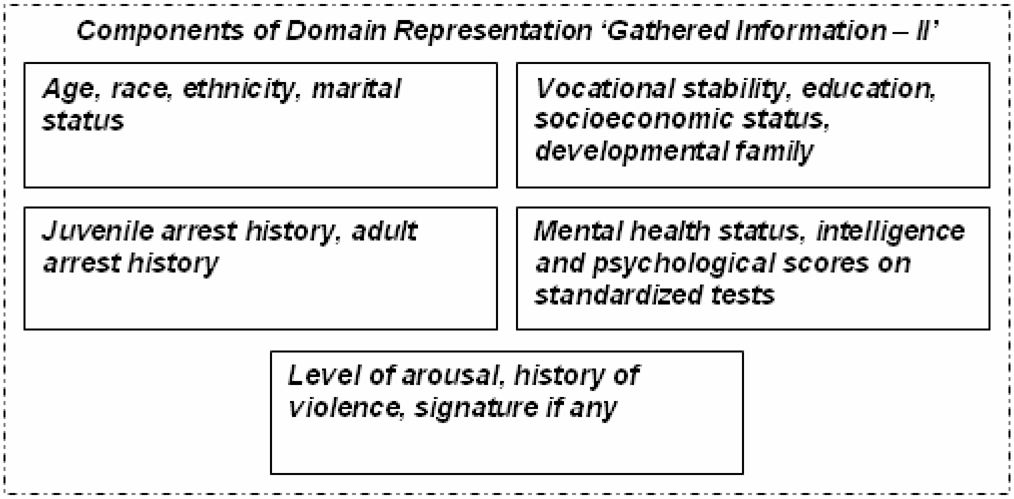
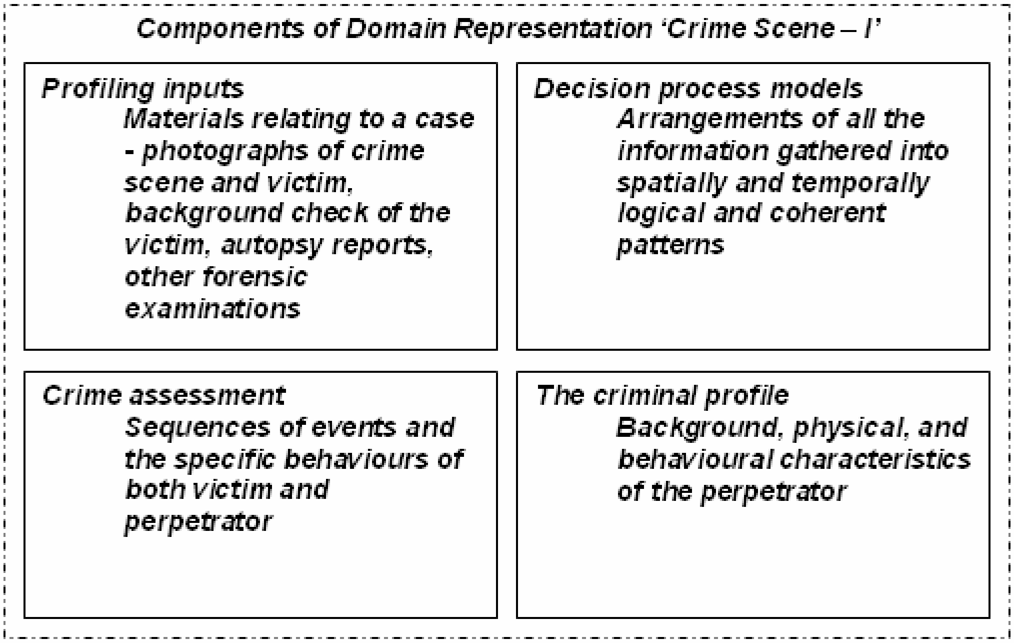

Components of Domain Representation 'Gathered Information – II'
Age, race, ethnicity, marital status
Vocational stability, education, socioeconomic status, developmental family
Juvenile arrest history, adult arrest history
Mental health status, intelligence and psychological scores on standardized tests
Level of arousal, history of violence, signature if any

Figure 4.6.: Domain Representation based on Gathered Information - II

4.4.1.3. Representation 3: Domain modeled on 'Crime Scene based Analysis – I'

Another perspective of modeling the domain comes from crime scene analysis. The following components are considered for this representation [Petherick, 2005]:

- Profiling inputs: Materials relating to a case, such as photographs taken of the crime scene and victim, a comprehensive background check of the victim, autopsy reports, other forensic examinations relating to the crime.
- Decision process models: Various arrangements of all the information gathered into spatially and temporally logical and coherent patterns.
- Crime assessment: Various sequences of events and the specific behaviours of both the victim and perpetrator - the roles of the victim and perpetrator.
- The criminal profile: Background, physical, and behavioural characteristics of the perpetrator.



Components of Domain Representation 'Crime Scene – I'

Profiling inputs
Materials relating to a case - photographs of crime scene and victim, background check of the victim, autopsy reports, other forensic examinations

Decision process models
Arrangements of all the information gathered into spatially and temporally logical and coherent patterns

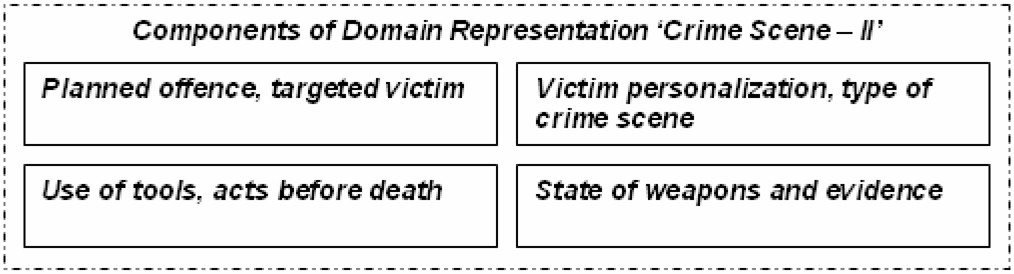
Crime assessment
Sequences of events and the specific behaviours of both victim and perpetrator

The criminal profile
Background, physical, and behavioural characteristics of the perpetrator

Figure 4.7.: Domain Representation based on Crime Scene - I

4.4.1.4. Representation 4: Domain modeled on 'Crime Scene based Analysis – II'

Another model of the domain from the crime scene perspective as above considers planned offence, targeted victim, victim personalization, type of crime scene, use of tools, acts before death, and state of weapons and evidence as its components. [Aked, 2005]



Components of Domain Representation 'Crime Scene – II'

Planned offence, targeted victim

Victim personalization, type of crime scene

Use of tools, acts before death

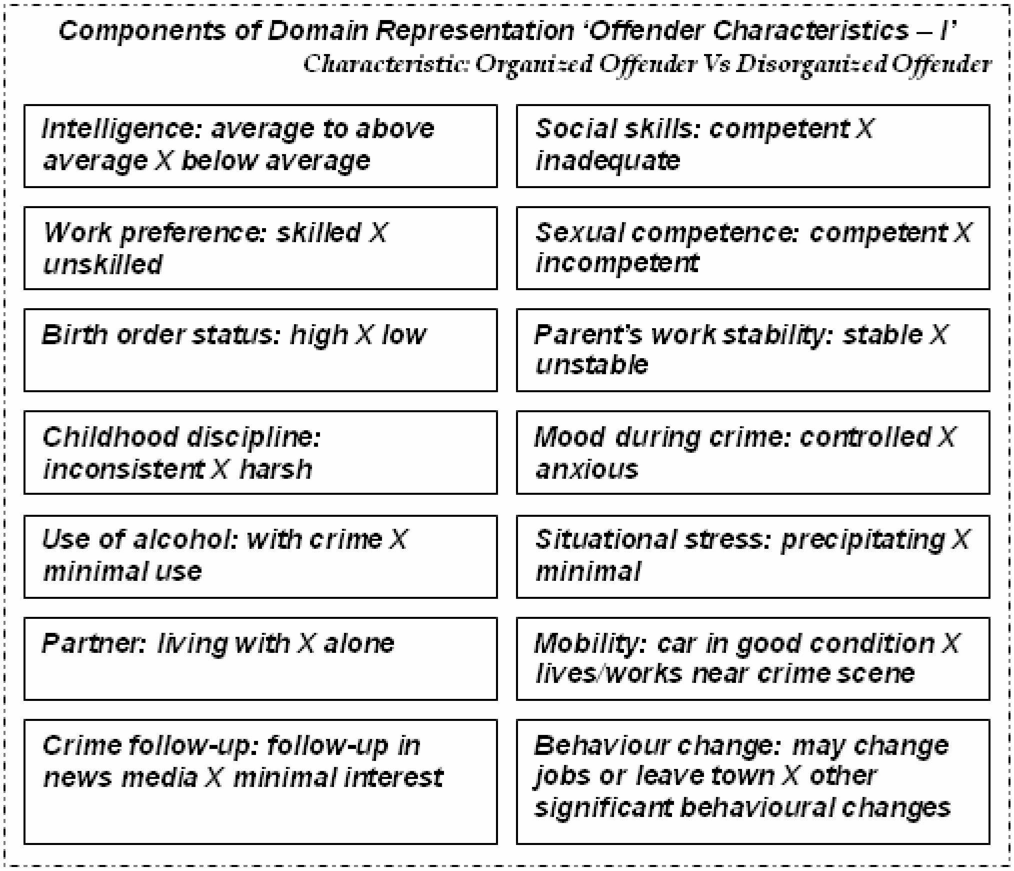
State of weapons and evidence

Figure 4.8.: Domain Representation based on Crime Scene - II

4.4.1.5. Representation 5: Domain modeled on 'Offender Characteristics based Analysis – I'

Criminology can be modeled from the perspective of the type of offender. This model is a finer representation of the domain. There are essentially two types of offenders: organized and disorganized. [Petherick, 2005][Aked, 2005] The components of the model that establish either one are:

- Characteristic: Organized **X** Disorganized
    - Intelligence: average to above average **X** below average
    - Social skills: competent **X** inadequate
    - Work preference: skilled **X** unskilled
    - Sexual competence: competent **X** incompetent
    - Birth order status: high **X** low
    - Parent's work stability: stable **X** unstable
    - Childhood discipline: inconsistent **X** harsh
    - Mood during crime: controlled **X** anxious
    - Use of alcohol: with crime **X** minimal use
    - Situational stress: precipitating **X** minimal
    - Partner: living with **X** alone
    - Mobility: car in good condition **X** lives/works near crime scene
    - Crime follow-up: follow-up in news media **X** minimal interest
    - Behaviour change: may change jobs or leave town **X** other significant behavioural changes



Components of Domain Representation 'Offender Characteristics – I'

Characteristic: Organized Offender Vs Disorganized Offender

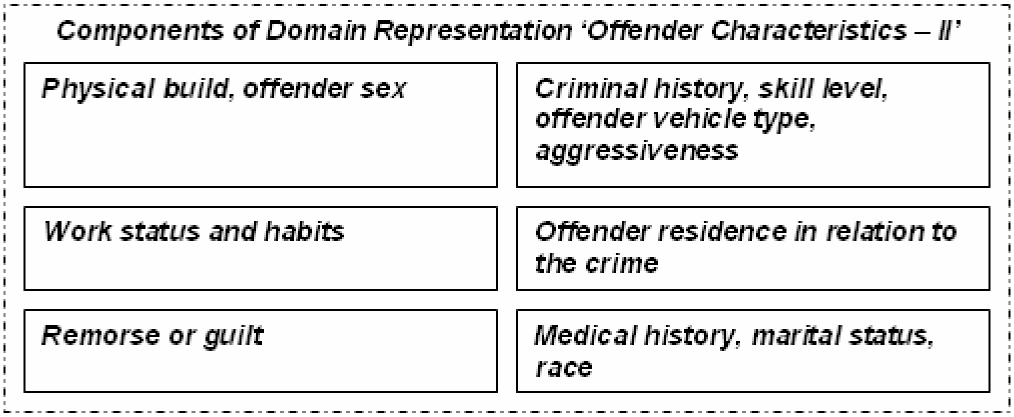
Intelligence: average to above average X below average	Social skills: competent X inadequate
Work preference: skilled X unskilled	Sexual competence: competent X incompetent
Birth order status: high X low	Parent's work stability: stable X unstable
Childhood discipline: inconsistent X harsh	Mood during crime: controlled X anxious
Use of alcohol: with crime X minimal use	Situational stress: precipitating X minimal
Partner: living with X alone	Mobility: car in good condition X lives/works near crime scene
Crime follow-up: follow-up in news media X minimal interest	Behaviour change: may change jobs or leave town X other significant behavioural changes

Figure 4.9.: Domain Representation based on Offender Characteristics - I

Such components are best considered as individual representational components rather than partial structures of any sort within the corpus

4.4.1.6. Representation 6: Domain modeled on 'Offender Characteristics based Analysis – II'

Another model of offender characteristics based analysis contains the following components. Note that the perspective of the domain that is being modeled here is the same as above, but the components are slightly different: physical build, offender sex, work status and habits, remorse or guilt, offender vehicle type, criminal history, skill level, aggressiveness, offender residence in relation to the crime, medical history, marital status, and race. [Petherick, 2005][Aked, 2005]



Components of Domain Representation 'Offender Characteristics – II'

Physical build, offender sex

Criminal history, skill level, offender vehicle type, aggressiveness

Work status and habits

Offender residence in relation to the crime

Remorse or guilt

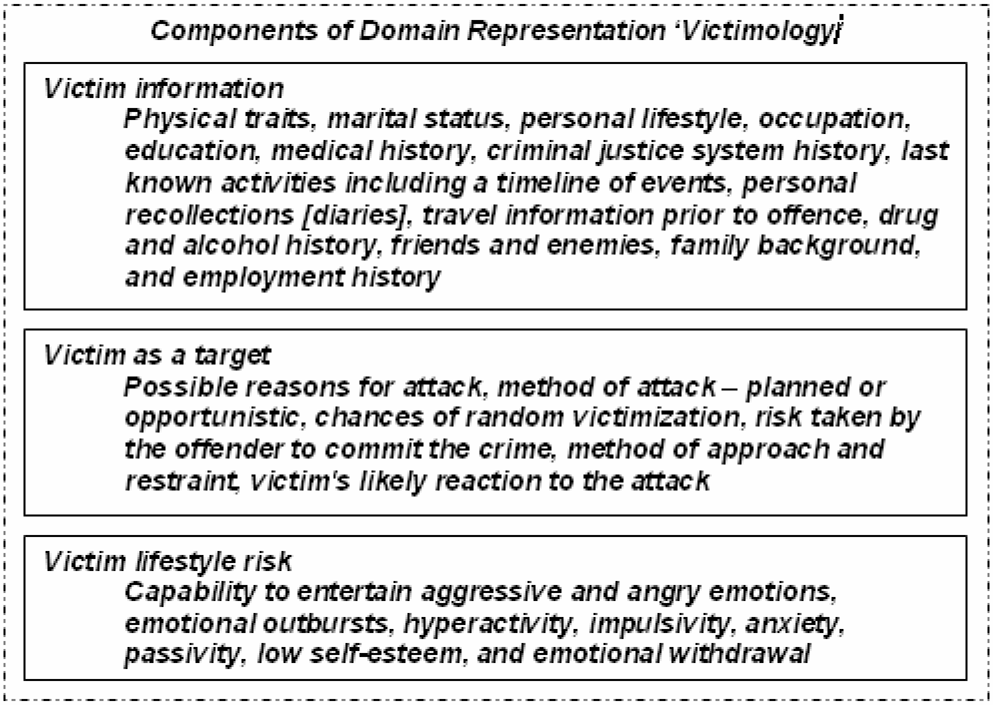
Medical history, marital status, race

Figure 4.10.: Domain Representation based on Offender Characteristics - II

4.4.1.7. Representation 7: Domain modeled on 'Victimology based Analysis'

Victimology documents the victims, their health and personal history, social habits and personality, and provides ideas as to why they were chosen as victims. In many situations, the offender will hold back from choosing a victim until he meets one that meets his needs based on fulfilling some fantasy or desire. This gives an insight into how the offender thinks and acts. By understanding how, where, when and why the victim is chosen, a relationship between the offender and victim can be established. The following components make up the representation of the domain from this perspective [Petherick, 2005]:

- Victim information: physical traits, marital status, personal lifestyle, occupation, education, medical history, criminal justice system history, last known activities including a timeline of events, personal recollections [diaries], travel information prior to offence, drug and alcohol history, friends and enemies, family background, and employment history.
- Victim as a target: possible reasons for attack, method of attack – planned or opportunistic, chances of random victimization, risk taken by the offender to commit the crime, method of approach and restraint, victim's likely reaction to the attack.
- Victim lifestyle risk: capability to entertain aggressive and angry emotions, emotional outbursts, hyperactivity, impulsivity, anxiety, passivity, low self-esteem, and emotional withdrawal.



Components of Domain Representation 'Victimology'

Victim information
Physical traits, marital status, personal lifestyle, occupation, education, medical history, criminal justice system history, last known activities including a timeline of events, personal recollections [diaries], travel information prior to offence, drug and alcohol history, friends and enemies, family background, and employment history

Victim as a target
Possible reasons for attack, method of attack – planned or opportunistic, chances of random victimization, risk taken by the offender to commit the crime, method of approach and restraint, victim's likely reaction to the attack

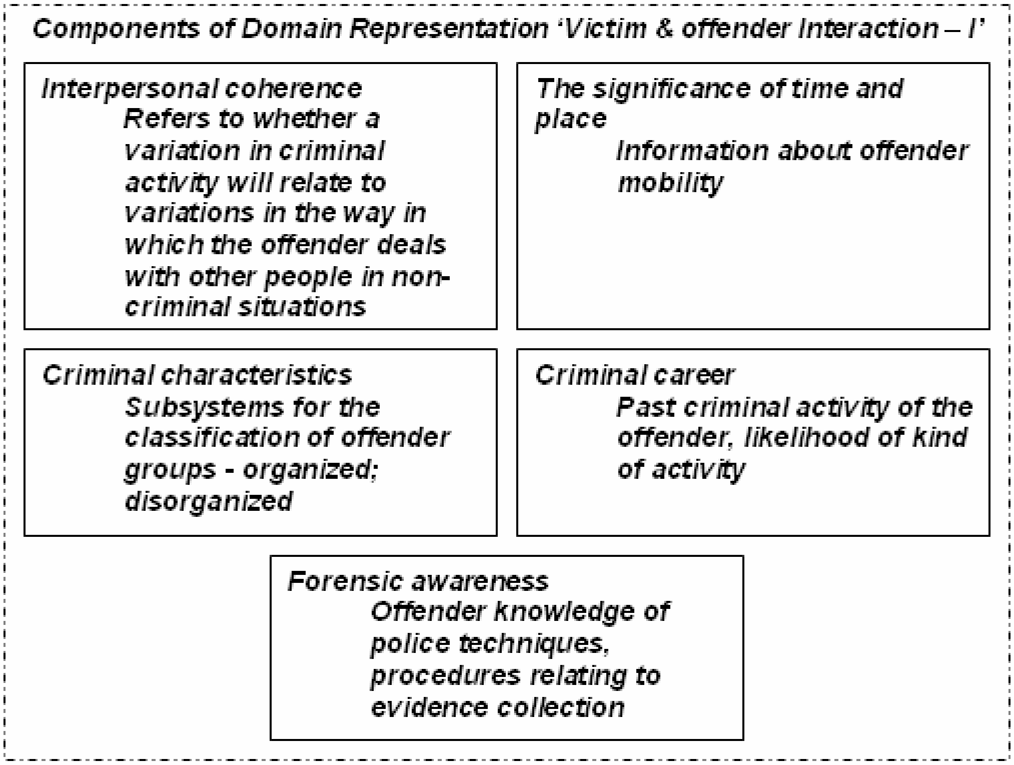
Victim lifestyle risk
Capability to entertain aggressive and angry emotions, emotional outbursts, hyperactivity, impulsivity, anxiety, passivity, low self-esteem, and emotional withdrawal

Figure 4.11.: Domain Representation based on Victimology

4.4.1.8. Representation 8: Domain modeled on 'Interaction between Victim and Offender – I'

This model of representing the Criminology domain is based on the interaction that takes place between the victim and the offender. The components of this representation include [Petherick, 2005]:

- Interpersonal coherence: Refers to whether a variation in criminal activity will relate to variations in the way in which the offender deals with other people in non-criminal situations.
- The significance of time and place: information about offender mobility.
- Criminal characteristics: subsystems for the classification of offender groups [e.g.: organized, disorganized].
- Criminal career: criminal activity engaged in by the offender in the past, what kind of activity this is most likely to have been.
- Forensic awareness: offender knowledge of police techniques and procedures relating to evidence collection, such as wearing of gloves, use of a condom, etc.



Components of Domain Representation 'Victim & offender Interaction – I'

Interpersonal coherence
Refers to whether a variation in criminal activity will relate to variations in the way in which the offender deals with other people in non-criminal situations

The significance of time and place
Information about offender mobility

Criminal characteristics
Subsystems for the classification of offender groups - organized; disorganized

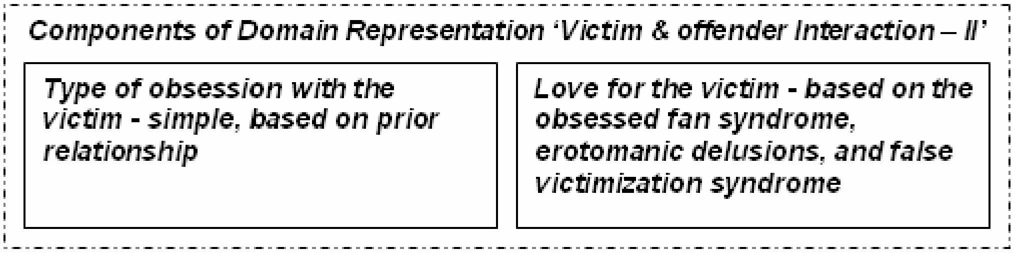
Criminal career
Past criminal activity of the offender, likelihood of kind of activity

Forensic awareness
Offender knowledge of police techniques, procedures relating to evidence collection

Figure 4.12.: Domain Representation based on Victim & Offender Interaction - I

4.4.1.9. Representation 9: Domain modeled on 'Interaction between Victim and Offender – II'

In another model, that represents the same perspective as above from a cyber stalking point of view, components such as type of obsession with the victim – simple, based on prior relationship, and love, based on the obsessed fan syndrome, erotomanic delusions, and false victimization syndrome are part of the offender-victim relationship. [Petherick, 2005]



Components of Domain Representation 'Victim & offender Interaction – II'

Type of obsession with the victim - simple, based on prior relationship

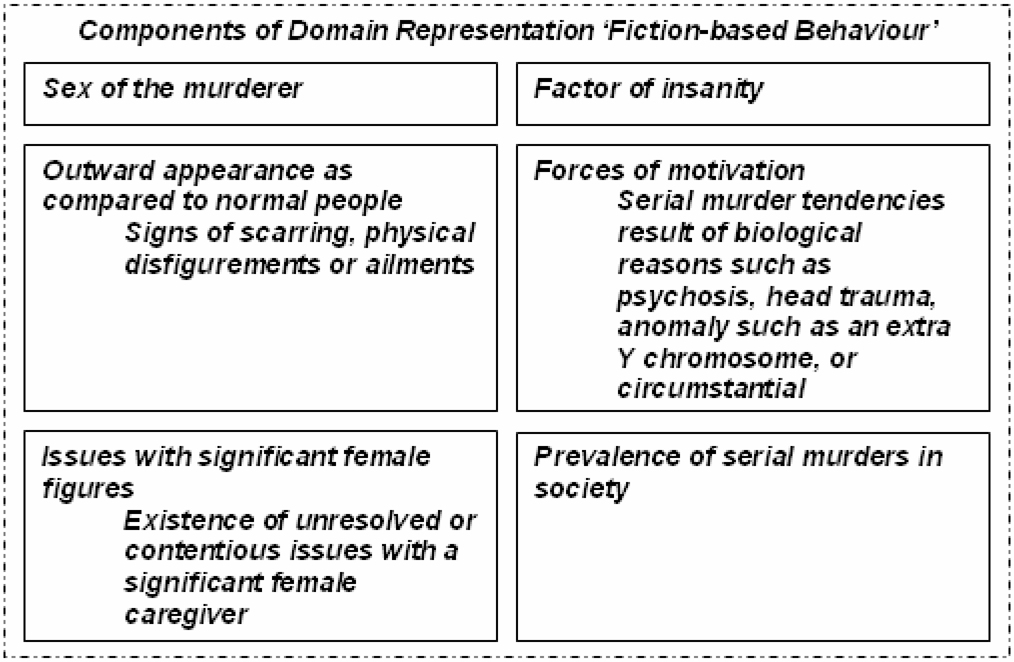
Love for the victim - based on the obsessed fan syndrome, erotomanic delusions, and false victimization syndrome

Figure 4.13.: Domain Representation based on Victim & Offender Interaction - II

4.4.1.10. Representation 10: Modeled on 'Fiction based Behavioural Analysis''

In the domain of Criminology, serial murder is best explained on the basis of the biopsychosocial model, in which it is assumed that one's developmental outcomes are largely the result of one's biology, or one's physical make up, one's psychology, or one's mental make up, and the social forces in one's life. No one influence is considered to be more important than the other and all three influences together determine the individual's overall personality and behaviour. [Petherick, 2005] One of the domain models of Criminology that represent the aspect of serial murderers, has been generated from the perspective of fiction and fantasy based on popular media influenced conceptions. The components that are part of this perspective of behavioural analysis include:

- Sex of the murderer: Usually male.
- Factor of insanity: Insanity in legal terms is the inability of an individual to know that their offending actions are ethically, morally, or socially unacceptable. [Petherick, 2005]
- Outward appearance as compared to normal people: Signs of scarring, physical disfigurements or ailments. Usually no particular distinguishing factor.
- Forces of motivation: Whether the serial murder is the result of biological reasons such as 'limbic psychosis', head trauma resulting from childhood accidents, or other biological anomaly such as an extra Y chromosome, or circumstantial.
- Issues with significant female figures: Existence of unresolved or contentious issues with a significant female caregiver.
- Prevalence of serial murders in society: Known that the US has approximately 75% of the world's known serial killers.



Components of Domain Representation 'Fiction-based Behaviour'

Sex of the murderer

Factor of insanity

Outward appearance as compared to normal people

Signs of scarring, physical disfigurements or ailments

Forces of motivation

Serial murder tendencies result of biological reasons such as psychosis, head trauma, anomaly such as an extra Y chromosome, or circumstantial

Issues with significant female figures

Existence of unresolved or contentious issues with a significant female caregiver

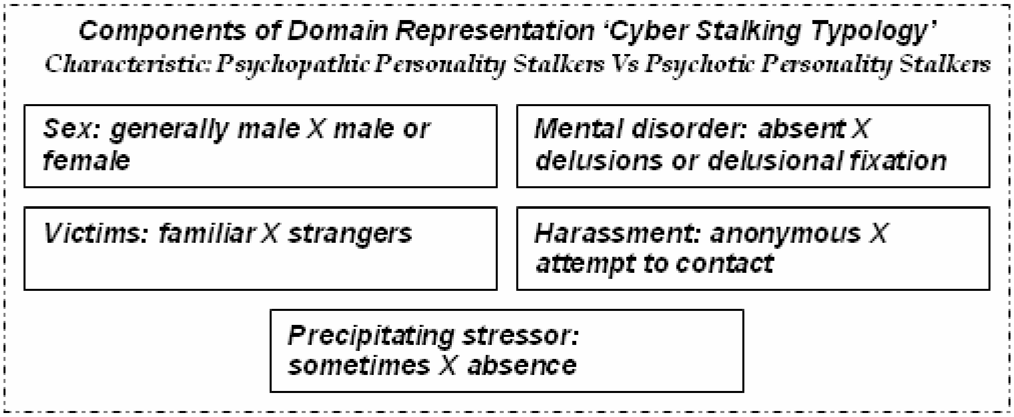
Prevalence of serial murders in society

Figure 4.14.: Domain Representation based on Fiction based Behaviour

4.4.1.11. Representation 11: Domain modeled on 'Cyber Stalking Typology'

Stalking occurs when an individual's behaviour is related to a thought, and possession of that thought is not enough, leading to practice of prohibited conduct. Stalkers can be classified broadly into two categories: psychopathic personality stalkers and psychotic personality stalkers. The following are the components that model this perspective of the domain. Again, this model is a finer, more atomic representation of the domain. [Petherick, 2005]

- Characteristic: psychopathic personality stalkers **X** psychotic personality stalkers
  - Sex: generally male **X** male or female
  - Mental disorder: absent **X** delusions or delusional fixation
  - Victims: familiar **X** strangers
  - Harassment: anonymous **X** attempt to contact
  - Precipitating stressor: sometimes **X** absence



Components of Domain Representation 'Cyber Stalking Typology'
Characteristic: Psychopathic Personality Stalkers Vs Psychotic Personality Stalkers

Sex: generally male X male or female

Mental disorder: absent X delusions or delusional fixation

Victims: familiar X strangers

Harassment: anonymous X attempt to contact

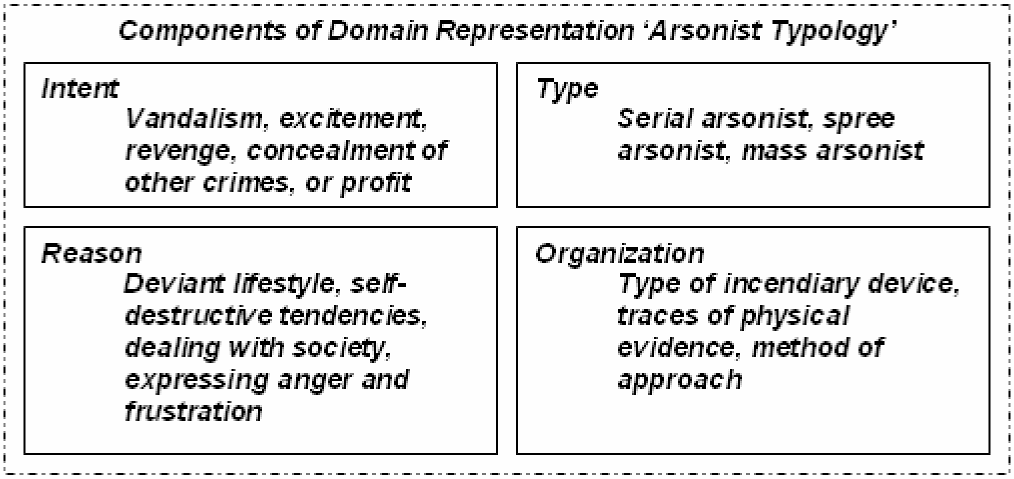
Precipitating stressor: sometimes X absence

Figure 4.15.: Domain Representation based on Victim Cyber Stalking Typology

4.4.1.12. Representation 12: Domain modeled on 'Arsonist Typology'

Any burning of property with malicious or self-gratifying intent can be defined as arson. The following components serve to model this perspective of the domain of Criminology [Aked, 2005]:

- Intent: vandalism, excitement, revenge, concealment of other crimes, or profit.
- Type: serial arsonist, spree arsonist, mass arsonist.
- Reason: deviant lifestyle, self-destructive tendencies, dealing with society, expressing anger and frustration.
- Organization: type of incendiary device, traces of physical evidence, method of approach.



Components of Domain Representation 'Arsonist Typology'

Intent
Vandalism, excitement, revenge, concealment of other crimes, or profit

Type
Serial arsonist, spree arsonist, mass arsonist

Reason
Deviant lifestyle, self-destructive tendencies, dealing with society, expressing anger and frustration

Organization
Type of incendiary device, traces of physical evidence, method of approach

Figure 4.16.: Domain Representation based on Arsonist Typology

4.5. Application of the CBKR+ Framework to the Domain of Criminology

Section 3.4 outlined some guidelines for application of the CBKR+ framework. Once the domain for the corpus and a set of truth values or belief coefficients, the confidence intervals and the probabilities for a given criterion have been established, these guidelines are followed as below:

- Identify the initial components, subsets, partial structures, and structures in the corpus, and establish the granularity best suited to the domain.
- Apply the Basis of Categorization model to the corpus to facilitate a first-level categorization of the corpus into possible associations and association makers. The associations are assigned a predetermined goodness value.
- Apply the Set of Criteria for Association to identify and extract more objective associations between the components of the corpus.
- Identify a hypothesis that has to be tested, and based on the associations formed between the components critical to the hypothesis, check if the hypothesis has been tested as true or false. Perform prediction in context by (a) observing the regularity of the occurrence of associations, (b) estimating the number of occurrences and ensuring that they fall within predetermined confidence intervals, (c) comparing this number to the predefined probabilities for associations and their impact on the state of the system.

The Levels of the Basis of Categorization model as described in the previous chapter are:

Level 0: This indicates the domain that must be represented.

Level 1: Association Value, stands for the goodness or the value of an association.

Level 2A: Domain Limits and Validity, stands for limits of the knowledge contained in the domain.

Level 2B: Component Nature, stands for the nature of being of the components of the domain.

Level 2C: Component Relationships, stands for the behavioral characteristics of the components.

The Criteria for Association to be applied to the corpus are the following. Subsets are groups of components of the representations that cannot discernibly stand alone, and partial structures are sections of the corpus – groups of associated components and subsets that can hold their own.

– Extraction of Roles: Who, What, When, Why, Where of the representational components.
– Component Key Co-occurrence: The components in the partial structure may occur in the same representation schema, in connection with another or separately. They may be associated with one another in the same representation schema or in different representation schemas.
– Key Subset Detection: Subsets or groups of components that are most associated with the key component.
– Component Attributes: Descriptive features that are most frequently associated with a component.
– Component Connectivity: Dependence between components and subsets relating to cause and effect.
– Component Collocations and Frequencies: Juxtapositions and arrangements that most frequently occur with respect to a given representational component, and the frequency with which they occur to form similar subsets.
– Search Activity & Association: Associations that arise between the components and subsets when a search is conducted for a given query, consecutively or in some definable pattern.
– Abstraction: Domain Subsets: Depending on some criteria of association, which components and subsets are related to one another and form partial structures of those components and subsets that adhere to that association.

4.5.1. Identifying the Components

Following the suggested guidelines, the components of each representation of the domain are shown as the individual blocks of every diagram covered in section 4.4. Within the corpus, the partial structures and structures are organized to some extent because the corpus was developed in a structured manner.

4.5.2. Applying the Basis of Categorization Model

The Basis of Categorization model is applied to the corpus in order to facilitate a first-level categorization of the corpus into possible associations, i.e., Level 2A and Level 2B, and association makers, i.e., Level 2C.

In the structures of Gathered Information – I and Gathered Information – II, <Survey report>, <Pictorial evidence available>, <Neighbourhood information>, <Forensic report available>, <Data about the victim>, <Age, race, ethnicity, marital status>, <Vocational stability, education, socioeconomic status, developmental family>, <Movements and activities of the victim> and <Juvenile arrest history, adult arrest history> are related to Levels 2A and 2B, while <Mental health status, intelligence and psychological scores on standardized tests> is related to Level 2C. Given the structures Crime Scene – I and Crime Scene – II, <Profiling inputs>, <Crime assessment>, <The criminal profile>, <Planned offence, targeted victim>, <Victim personalization, type of crime scene>, <Use of tools, acts before death> and <State of weapons and evidence> conform to Levels 2A and 2B, while <Decision process models> conforms to Level 2C. In Offender Characteristics – I and Offender Characteristics – II <Intelligence>, <Social skills>, <Work preference>, <Sexual competence>, <Birth order status>, <Parent's work stability>, <Childhood discipline>, <Mood during crime>, <Use of alcohol>, <Situational stress>, <Partner>, <Mobility>, <Crime follow-up>, <Behaviour change>, <Physical build, offender sex>, <Criminal history, skill level, offender vehicle type, aggressiveness, Offender residence in relation to the crime>, <Work status and habits> and <Medical history, marital status, race> conform to Level 2A and Level 2B, while <Remorse or guilt> falls in Level 2C. From Victimology, <Victim information>, <Victim as a target> and <Victim lifestyle risk> fall in Levels 2A and 2B. In the structures Victim & Offender Interaction – I and Victim & Offender Interaction – II, <The significance of time and place>, <Criminal characteristics>, <Criminal career>, <Forensic awareness> and <Type of obsession with the victim> relate to Levels 2A and 2B, while <Interpersonal coherence> and <Love for the victim> relate to Level 2C. On similar principles, from Fiction-based Behaviour, <Sex of the murderer>, <Factor of insanity>, <Outward appearance as compared to normal people> and <Issues with significant female figures> fall in 2A and 2B, while <Forces of motivation> and <Prevalence of serial murders in society> fall in 2C. In Cyber Stalking Typology, components <Sex>, <Victims>, <Mental disorder> and <Harassment> relate to 2A and 2B, whereas <Precipitating stressor> relates to 2C. Finally, in Arsonist Typology, components <Type> and <Organization> fall in 2A and 2B, while components <Intent> and <Reason> fall in 2C. This segregation is shown in the following Table 4.1.

Possible Associations		***Association Makers***
Level 2A: Domain Limits and Validity	***Level 2B: Component Nature***	***Level 2C: Component Relationships***
Gathered Information – I and Gathered Information – II		
<Survey report>, <Pictorial evidence available>, <Neighbourhood information>, <Forensic report available>, <Data about the victim>, <Age, race, ethnicity, marital status>, <Vocational stability, education, socioeconomic status, developmental family>, <Juvenile arrest history, adult arrest history>, <Movements and activities of the victim>		<Mental health status, intelligence and psychological scores on standardized tests>
Crime Scene – I and Crime Scene – II		
<Profiling inputs>, <Crime assessment>, <The criminal profile>, <Planned offence, targeted victim>, <Victim personalization, type of crime scene>, <Use of tools, acts before death>, <State of weapons and evidence>		<Decision process models>
Offender Characteristics – I and Offender Characteristics – II		
<Intelligence>, <Social skills>, <Work preference>, <Sexual competence>, <Birth order status>, <Parent's work stability>, <Childhood discipline>, <Mood during crime>, <Use of alcohol>, <Situational stress>, <Partner>, <Mobility>, <Crime follow-up>, <Behaviour change>, <Physical build, offender sex>, <Criminal history, skill level, offender vehicle type, aggressiveness, Offender residence in relation to the crime>, <Medical history, marital status, race>, <Work status and habits>		<Remorse or guilt>
Victimology		
<Victim information>, <Victim as a target>, <Victim lifestyle risk>		
Victim & Offender Interaction – I and Victim & Offender Interaction – II		
<The significance of time and place>, <Criminal		<Interpersonal coherence>

characteristics>, <Criminal career>, <Forensic awareness>, <Type of obsession with the victim>	and <Love for the victim>
Fiction-based Behaviour	
<Sex of the murderer>, <Factor of insanity>, <Outward appearance as compared to normal people>, <Issues with significant female figures>	<Forces of motivation> and <Prevalence of serial murders in society>
Cyber Stalking Typology	
<Sex>, <Victims>, <Mental disorder>, <Harassment>	<Precipitating stressor>
Arsonist Typology	
<Type> and <Organization>	<Intent>, <Reason>

Table 4.1.: First-level Categorization of the Components in the Corpus

The first-level categorization done, Level 1 ascribes a predetermined goodness value to the associations that exist between the components, and becomes useful to hypothesis testing and prediction in context after the application of the Set of Criteria for Association. Now, the application of Set of Criteria for Association to this corpus shall identify more objective associations between the components.

4.5.3. Applying the Set of Criteria for Association

The Set of Criteria for Association identifies and extracts more objective associations in the corpus. Using each criterion as briefly recounted above:

– *Extraction of Roles*: Using Extraction of Roles, each component in every representation has a descriptive who, what, when, why and where value. The associativity for each, i.e. values for associated who(s), values for associated what(s), and so on can serve to create the related partial structures in the corpus. To examine this in greater detail, the five W(s) are applied to the subset of <Survey report> from the partial structure General Information – I and <Reason> from Arsonist Typology. <Survey report> is the subset name, and is a subset or partial structure within the corpus. It shall be active for its entire lifetime as it contains information about the crime in a survey report of the entities involved and the flow of events form. This partial structure is part of the structure or subset of Gathered Information – I. <Reason> is the component name and it is part of the partial structure of Arsonist Typology. It shall be also active for its entire lifetime as it describes the various reasons for a perpetrator to commit arson. This component is a part of the structure Arsonist Typology.

- *Component Key Co-occurrence*: For Component Key Co-occurrence, the components in the subsets of <Decision Process Model> from Crime Scene – I, <The Criminal Profile> from Crime Scene – I, and the components of <Intelligence>, <Work preference>, <Birth order status>, <Childhood discipline>, <Use of alcohol>, <Partner>, <Social skills>, <Sexual competence>, <Parent's work stability>, <Mood during crime>, <Situation stress>, and <Behaviour change> from Offender Characteristics – I occur in conjunction with <Profiling inputs> from Crime Scene – I. Similarly, the components <Types of obsession with the victim> and <Love for the victim> from Victim and Offender Interaction – II, <Intent> from Arsonist Typology, and <Mental Health et al> from Gathered Information – II occur in the same situational instance as <Forces of motivation> from Fiction based Behaviour. Since certain components in the partial structure occur in connection with another or separately, they form the association of key co-occurrence that relates them through the property that these components are found to occur together.
- *Key Subset Detection*: In Key Subset Detection, for the key component of <Forensic awareness> from Victim and offender Interaction – I, the subsets of <Use of tools et al> and <State of weapons et al> from Crime Scene – II, <Victim as target> from Victimology, <Organization> from Arsonist Typology, and <Criminal history et al> from Offender Characteristics – II are the key subsets associated with the key component. These subsets or groups of components are now related through the criterion that they are most associated with the key component of <Forensic awareness>.
- *Component Attributes*: For the adjective qualities of Component Attributes, <Victim information> from Victimology and <Data about the victim> from Gathered Information – I are all descriptive in nature and serve to create a picture of the victim. Similarly, <Pictorial evidence available> and <Forensic report available> from Gathered Information – I, <Crime assessment> from Crime Scene – I, and the component <Signature if any> from Gathered Information – II, all contribute to describe whether the crime has a signature or

not, and if it does what kind. The relation is of the description components that are most associated with a given component.

- *Component Connectivity*: Using Component Connectivity, from Crime Scene – I, <Crime assessment> is dependent on <Decision process models>, which is a product of <Profiling inputs>. Therefore, <Profiling inputs> → cause → <Decision process models> → effect → <Crime assessment>. Similarly, the subsets of Gathered Information – II may cause the components of Offender Characteristics – I, as well as those of Victimology. In another instance, the <Victim information>, <Victim as a target>, <Victim lifestyle risk> from Victimology and the components of Offender Characteristics – I and Offender Characteristics – II may cause the subsets of Victim and offender Interaction – I and Victim and offender Interaction – II. Also, the components of Victim and offender Interaction – II affect the components of Cyber Stalking Typology. Establishing the cause and effect of components is a result of knowledge gained through study of the domain of Criminology from the sources referenced in this chapter.
- *Component Collocations and Frequencies*: For Component Collocations and Frequencies, given subset <Survey report> from Gathered Information – I, all subsets and components from Fiction based Behaviour, and Cyber Stalking Typology, the different arrangements and juxtapositions are best represented by the component of <Decision process models> from Crime Scene – I.
- *Search Activity and Association*: Using Search Activity and Association, given consecutive searches for age of offender, sex of offender (from Offender Characteristics – I), age and sex of victim, and conditions for victim as target (from Victimology), and factor of insanity of perpetrator (from Fiction based Behaviour), with results of male, familiar victim, no mental disorder, and anonymous harassment, it may be a psychopathic stalker who may be an arsonist.
- *Abstraction: Domain Subsets*: Abstraction: Domain Subsets yields domain points around Offender in the form of the subsets or partial structures of Offender Characteristics – I and Offender Characteristics – II, around Victim in the form of subsets from Victimology and <Data about the Victim> from Gathered Information – I, around the Association of both in the form of the subsets of

Victim and offender Interaction – I and Victim and offender Interaction – II, and around Crime in the form of <Intent> and <Reason> from Arsonist Typology.

4.5.4. Hypothesis Testing and Prediction in Context

The two additional applications of hypothesis testing and prediction in context that the CBKR+ framework adds to the repertoire of the CBKR methodology are discussed as follows. On the collated corpus of representational components, within given constraints, hypotheses such as the following can be tested:

- "given <juvenile arrest history> from Gathered Information – II, the <intent> of vandalism from Arsonist Typology, and <forensic awareness> from Victim and offender Interaction – I, the arsonist organized the crime using incendiary devices that cannot be traced and leave no physical evidence (Arsonist Typology)",
- "given <pictorial evidence of signature> from Gathered Information – I, and <known signature type> from Gathered Information – II, the perpetrator was a serial murderer (Fiction based Behaviour) preying on a victim satisfying his needs (Victimology)",
- "given <Victim lifestyle risk> from Victimology, <The significance of time and place> from Victim & offender Interaction – I, and <Movements and activities of the victim> from Gathered Information – I, the perpetrator is likely to have planned an offence against targeted victim (Crime Scene – I)", and
- "given <Interpersonal coherence> from Victim & offender Interaction – I, <Intelligence and psychological scores on standardized tests> from Gathered Information – II, and <Outward appearance as compared to normal people> from Fiction-based Behaviour, the perpetrator is likely to harass the victim through cyberspace (Cyber Stalking Typology)".

Similarly, given the requisite confidence intervals or degrees of belief, prediction in context can take the form of:

- "given appropriate associations between <extra Y chromosome> and <issues with significant female figures> from Fiction based Behaviour, <forensic awareness> from Victim and offender Interaction – I and <anonymous harassment> from Cyber Stalking Typology, it can be predicted that such perpetrators should be under surveillance for cyber stalking with a propensity for serial murder."

- "given appropriate associations between the components of Offender Characteristics – I and <juvenile arrest history> from Gathered Information – II, a likelihood of disorganized criminal activity re-occurrence can be predicted."
- "given appropriate associations between <Use of alcohol> from Offender Characteristics – I, <Criminal career> from Victim & offender Interaction – I, and <Issues with significant female figures> from Fiction-based Behaviour, a propensity for attacking or harming women when intoxicated can be predicted."
- "given appropriate associations between <Parent's work stability> from Offender Characteristics – I, <Socioeconomic status> from Gathered Information – II, and <Criminal career> from Victim & offender Interaction – I, it can be predicted that perpetrators from instable, low-income economic background are likely to adopt a criminal career."

4.6. The CBKR+ Framework vis-à-vis the CBKR Methodology

Sections 4.3 and 4.5 demonstrate the application of the CBKR+ framework to examples from the CBKR methodology and to the real world domain of Criminology. In the CBKR examples in section 4.3, it is seen that the CBKR methodology is unable to derive specific associations between the components. For factual and similarity queries posed on the corpus, a cluster of components is returned, but the methodology is unable to say how the components are specifically associated. When the corpus is analyzed using the CBKR+ framework, it is seen that specific associations between the components are formed using the Basis of Categorization model and the Set of Criteria for Association. The CBKR+ framework identified and extracted specific associations, such as, <GPA, StudentID, value, address>, <Car-Review, LuxuryClass, VeryGood, goodTires>, and <Car-Review, LuxuryClass, Lexus, Toyota>.

The enhanced ability of the CBKR+ framework over the CBKR methodology is better shown through the application of the framework to the domain of Criminology, as follows:

- *Improved Association Capability*: The CBKR methodology is limited to addressing factual and similarity queries over a given corpus. As a result, the methodology would only return a set of components that were clustered together by a binding factor. For a query posed about victims, components that modeled all victim information would be returned, such as, <Victim information>, <Victim as a target>, and <Victim lifestyle risk> from Victimology, and the components of Victim & Offender Interaction – I and Victim & Offender Interaction – II. The methodology would be unable to associate <Data about the victim> from Gathered Information – I to give a more complete representation of the victim. Using CBKR+:

  - First, the relevant components of the three structures are categorized as possible associations by the Basis of Categorization model, as is the component <Data about the victim> from Gathered Information – I.
  - Second, the criterion of Component Attributes also establishes an objective association between <Victim information> from Victimology and <Data about the victim> from Gathered Information – I.

  Similarly, the criterion of Search Activity and Association uses related specific searches for <Age of offender>, and <Sex of offender> from Offender Characteristics – I, <Age and sex of victim>, and <Conditions for victim as target> from Victimology, and <Factor of insanity of perpetrator> from Fiction based Behaviour to associate these components together, since most searches done for these components follow a similar pattern.

- *Hypothesis Testing*: This association can result in the hypothesis that a male, familiar with victim, having no mental disorder, and practicing anonymous harassment may be a psychopathic stalker who may be an arsonist. This hypothesis can be accepted or rejected based on the study of the domain and the regularity of occurrence of such a representation. If accepted, the components associated by the hypothesis and the associations themselves are integrated into the corpus. In contrast, the CBKR methodology is unable to form the association and leverage it as the CBKR+ framework does.
- *Prediction in Context*: Due to the enhanced association capability of the CBKR+ framework over the CBKR methodology, hypotheses, such as the ones outlined in section 4.5.4, can be tested given the necessary constraints. In addition, given the requisite confidence intervals or degrees of belief, prediction in context can take the form of the predictions outlined in the same section. As seen from examples throughout this chapter, the ability of prediction is better in the sense

that CBKR+ predicts the same things with less evidence. This is also seen in the example in section 4.2 through Figures 4.2 and 4.3.

For any socially applicable technology to be truly useful, it does not seem enough to pose simple queries. Firstly, there are many approaches that already work on the simple query principle, and secondly, with the multitude of data available, a technology that can manipulate it has more use than one that cannot. As regards the corpus based approach to representation, there is a need to be able to analyze the schemas in the corpus, and establish associations between their components. This is done to create a more complete representation of the domain than any one schema alone can. For example, going back to the example of books in a library in section 1.3, the ethos of a country includes the country's culture, attire, cuisine, art, heritage, customs, and so on. It is more efficient to segregate them for efficient organization, and yet have some association between these factors so that the term 'ethos' includes all these aspects. The CBKR methodology can collect all these aspects together, and present result sets to general queries made about the ethos country. The CBKR+ framework fine-tunes the association between these aspects – it segregates them, but keeps them connected through identified associations. This fine-tuning of the corpus increases the expressiveness and analysis capacity of the approach, thereby making it useful for applications, such as, hypotheses testing and prediction in context. The CBKR+ framework is a good indicator to study and outline the development of the domain and the completeness of its representation.

4.7. Chapter Summary

This chapter discussed the application of the proposed CBKR+ framework, and the guidelines suggested for it. The chapter illustrated the examples of the CBKR methodology, and indicated how the CBKR+ framework could enhance the capability of the methodology. The CBKR+ framework was then applied to the physical world domain of Criminology. The use of the framework to support the applications of Hypothesis Testing and Prediction in Context was demonstrated.

Chapter 5 – Conclusions and Future Work

To proceed from one truth to another, and connect distant propositions
by regular consequences, is the great prerogative of man.
~ Samuel Johnson (1709-1784), Writer

5.0. Chapter Introduction

This chapter presents the conclusions of the work in this research. Some limitations and biases of the proposed CBKR+ framework have been presented. Directions for future work have been outlined as well.

The chapter begins with a summary of the thesis, followed by a discussion of the advantages, biases and limitations of the proposed CBKR+ framework, and the relevance of the research contained in this work. This is followed by some directions for future research. The next section states the contributions of this work. The chapter ends with some conclusive comments about the thesis.

5.1. Thesis Summary

In the first chapter of this thesis, general representational systems were discussed. Though these traditional systems were based on principles of reason and intuition, their coding was time-consuming and difficult. They used complex evaluations as quantitative measure to judge the quality of specification of the representation and the resultant data. Enhancement of traditional representational systems has been largely geared towards very constrained queries on rule-based, absolute systems of knowledge based on mathematics and inventories. [Waltz & Kasif, 1995] Some criteria suggested by Torsun [Torsun, 1995] were recommended for the creation and evaluation of any form of a knowledge representation system. It was seen that one of the bigger problems in the area was that any representation of a given domain was inherently incomplete, i.e., it was almost impossible to represent a domain in its entirety. The researcher found that the Corpus Based Knowledge Representation [CBKR] methodology tried to alleviate this problem by using a corpus of representations collectively in order to arrive at a more complete representation of a given domain. The chapter included the research problem, and an intended solution to be arrived at via this thesis.

A brief study of traditional representation systems, as seen in chapter 2, established the fact that most traditional representation techniques were complex and could not fully represent a domain. The various methods of representation demonstrated that modeling a given domain and leveraging it for use in applications could be a costly, complex, and tedious challenge. This chapter also reviewed the CBKR methodology, a wholly new approach to knowledge representation, and the related research that led to this methodology. The approach complemented traditional representation schemes, and added to the repertoire of methods available to model a given domain and use it for various purposes. In contrast to traditional knowledge representation schemes, corpus-based approaches used the notion of a corpus, a collection of schema representations of a domain, to allow multiple representations of the domain to be used for query processing (e.g. factual and similarity queries). The need for matching and mapping between representational components of the domain was explained, along with the many efforts that have been employed to accomplish it. It was concluded that, although corpus-based approaches had been successfully applied in fields such as language translation and information retrieval, they were limited in terms of association capability and their support for hypothesis testing and prediction when applied to a corpus of schema representations of a given domain. The lack of association capability came from the fact that the methodology did not have criteria in place to extract substantial associations in the corpus. The lack of support for hypothesis testing and prediction in context, limited what could be analyzed using the CBKR approach to query-processing, without support for analysis and inference from the corpus. The discussion of the CBKR methodology was followed by a description of the Domains of Reasoning model. This provided the background information needed for the framework proposed by this research.

Chapter 3 introduced and discussed the proposed CBKR+ framework in detail. As part of this framework, the Basis of Categorization model and the Set of Criteria for Association were discussed as well. The framework increased the expressiveness of the corpus-based approach by identifying and incorporating association criteria to allow the support of new forms of analyses related to hypothesis testing and prediction in context that could be applied to domains such as criminology. This framework revealed a variety of properties of the domain, and helped to improve the completeness and expressiveness of the domain's representation. The relation between the Basis of Categorization model and the Set of Criteria for Association was discussed as well. The Set of Criteria for Association could not be a random collection of criteria if the set was to be effectively applied to a corpus containing the representational components of an entire domain. These criteria were to have some unifying factor so that they could work together, in conjunction, to facilitate the identification of associations between the components of the corpus. This unifying factor was the Basis of Categorization model. The working of the CBKR+ framework was discussed along with some guidelines to employ the framework. Finally, the proposed framework was discussed on the basis of a set of performance criteria suggested by Torsun [Torsun, 1995 as in Pesonen, 2002].

This was followed by the application of the CBKR+ framework to case studies in chapter 4. First, it was demonstrated on examples from the CBKR research. Then, the framework was applied to the physical world domain of Criminology, and included 12 domain representations acquired from multiple sources. In addition to the case study from the CBKR methodology in section 4.3, the application case study in sections 4.4 and 4.5 showed that the new framework could be applied to a wider spectrum of applications, such as hypothesis testing and prediction in context that leverage diagnostic and probability-based queries.

The researcher arrived at the conclusion that while CBKR was a competent methodology for representing and using knowledge of the form of domain representations, the proposed CBKR+ framework enhanced its competence in an elegant and easy to comprehend manner, as evidenced by the case studies above. The quality of elegance was viewed in terms of the simplicity and structure of the application of the framework. The general assets of the proposed framework were found to be that CBKR+ offered an enhancement to CBKR, in that, (1) fine-tuning of the corpus and improvement of the expressiveness of the domain representation by identification and extraction of associations between the components of the corpus was achieved, and (2) additional applications – hypothesis testing & prediction in context – based on analysis of the corpus could be addressed. The corpus's improved associative capacity and expressive adequacy could be applied to diagnostic and probability based applications such as hypothesis testing and prediction in context. There was improved flexibility in the CBKR+ framework as the proposed part-inductive and part-deductive association finding process allowed for a better organization of the schema components in the corpus, and made the addition of new data more structured. This research suggests that the enhancements proposed to CBKR be used in conjunction with the methodology, and not instead of it.

5.2. Biases and Limitations of the Proposed Approach

One of the biases of this research is that the proposed framework is theoretical work, based on observed real world domains. The CBKR+ approach works for large domains where patterns in data can be identified. Representational models of demographic and ethnographic domains such as environmental monitoring, marketing strategies, and criminology have a higher possibility of containing potential patterns. A practical implementation of the CBKR+ framework was impossible due to the difficulty in procuring real data in the form of actively employed schema representations for any of the above domains, due to the limitations of cost, availability, and time considerations. Domain representations were needed that were sufficiently complex, made up of a host of multivariate schemas, easily available, sound enough to be used for demonstration, and cost-effective. Such a single domain for environmental monitoring and marketing strategies was impossible to procure despite many endeavors. The common problems with acquiring real data for criminology were (1) the reluctance to share such data due to security concerns, and 2) non-availability of such data in soft format. The working of the framework is demonstrated through its application to the real world domain of Criminology. The researcher believes that in the absence of real schemas for any of the listed domains, demonstration of the CBKR+ framework to the domain of Criminology is likely to have the most social significance and use. This is because the domain of Criminology was found to have the least number of techniques available for representation. The researcher's intention was to contribute to a developing domain, rather than contribute to domains that had more mechanisms for representation. An implementation of the CBKR+ framework to Criminology, attempted with the limited amount of information that was available, would have given results not truly reflecting the capacity and usability of the model. CBKR+, essentially, is a conceptual framework. The working is therefore demonstrated through the case study on criminology that has been done using theoretical models from academic research.

Some of the other biases that the CBKR+ framework is subject to are:

- The framework appeals only to a certain class of applications, even though it extends the spectrum of applications the CBKR methodology could address. The CBKR methodology has been successfully applied in fields such as language translation and information retrieval, but it is limited in terms of association capability and its support for hypothesis testing and prediction in context. The proposed CBKR+ framework improves the association capability of CBKR, and can be used for applications, such as hypothesis testing and prediction in context, which are based on analyses of the corpus.
- The approach is applied to a restricted level of granularity of domain representation. The granularity cannot be very fine, as this may cause unnecessary and overlapping associations to be formed, thereby decreasing the simplicity of the approach. The components of the corpus represent a macroscopic view of the domain. As a result, this work identifies and extracts macroscopic associations. A microscopic view of the domain would be one where every entity and element within the domain, and not just its structure, would be modeled.
- As in any other case, the collated corpus is brittle in terms of the completeness of the domain representation. Even though the corpus has multiple representations, acquired from multiple sources, and in multiple formats, the domain representation is only as complete as that possible by the experts contributing to the corpus.
- The reasoning methodology gives results of gradual rather than dogmatic and highly precise correctness. Though this is desired in the CBKR and CBKR+ approaches, it is a bias, in that it makes these approaches applicable to a certain class of domains and applications. A property of such applications is that they are more domain schema reliant rather than domain data reliant, in that, leveraging the schemas can be more useful to acquire possible result sets rather than leveraging data to acquire exact, precise results.
- The identification and extraction of associations in the corpus is based on the regularity of occurrence of representational components, and the patterns

observed therein. Hypothesis testing and prediction in context are also based on patterns that are identified on the regularity of occurrence of the components in the corpus and the relations that exist between them.

- The approach works for large domains in which there is a scope of identifying patterns in the data. Representational models of demographic and ethnographic domains such as environmental monitoring, marketing strategies, and criminology have a higher possibility of containing potential patterns. This is because there is the possibility of (1) a large number of schema representations created by a variety of experts due to the large size of such domains, (2) similar data content that is modeled differently through many schemas giving rise to the formation of a corpus, and (3) the use of such domains to acquire approximate or partially correct result sets rather than strictly accurate ones, and (4) the use of such domains for multiple applications that are based on similar representations of the domain.

Any framework works well within the limitations set by the designer. The CBKR+ framework also has a set of limitations influencing it, and works best within them. These limitations are:

- An implementation of the model as applied to a real world domain is needed to fully understand the relative capabilities and limitations of the model outlined above. The greater goal, essentially, is to create a methodology for adaptive, complete, application independent representations of domains without rigid input methods and costly labor of human experts and programmers.
- In the physical world, hermetically sealed domains rarely exist. Most domains are co-dependent with other domains. The proposed framework reduces the incompleteness of the representation of a given domain. It however, does not alleviate the problem of cross-domain association. Associations may be established across different domains under the condition that the different domains must have some common underlying characteristics.
- The element of surprise [Gaines, 1987] is not considered. Dealing with unaccounted for or random, unexpected representational components and associations requires a more fuzzy form of logic. It is important to emphasize that the identification of associations in the proposed framework is based on

observing and quantifying the regularity of occurrence of representational components and the associations between them.

- The outcomes of the hypothesis testing and prediction in context can be false positives or false negatives. Such outcomes are not addressed by this work as they require a more fuzzy logic-like approach for resolution.

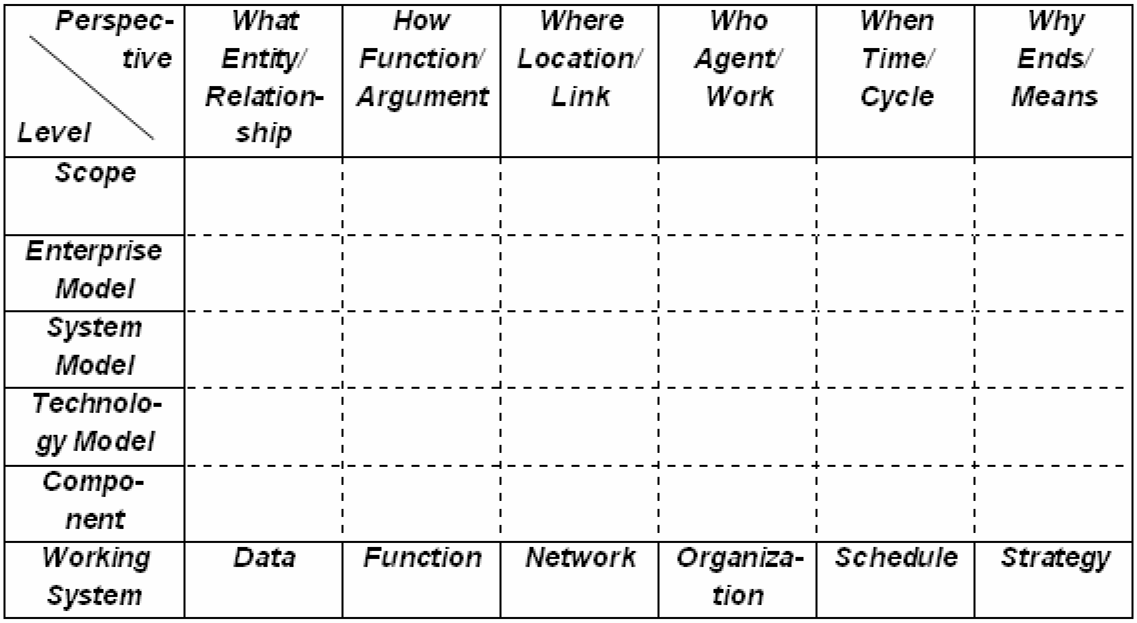
5.3. Directions for Future Work

The directions suggested for future research with the CBKR+ framework are those outlined below:

- The CBKR+ framework can be formalized using Formal Concept Analysis [Ganter & Wille, 1999, Faid et al, 1999]. Formalization of the framework will create a set of rules, grammars and mathematical logic for the associations formed that will be universally recognized and be easier to manipulate collectively. FCA has the capability to do this, and in addition, it can produce graphical visualizations of these associations based on the rules and logic it helps to create. These visualizations of formalized associations can be used to analyze the range and consistency of the relations.
- Formalizing the CBKR+ framework can be followed by creating a standard for comprehensive, multi-schema domain model representation – an XML-like set of rules that are universally applicable and understood.
- The research has proposed a preliminary Set of Criteria for Association to enhance the CBKR methodology in terms of fine-tuning the corpus of representational schema components and addressing a wider range of applications. With further research, more criteria can be added to enhance the CBKR+ framework to further refine the corpus and make the approach useful for even more types of applications.
- An immediate direction would be to obtain implemented representational schemas for a domain and apply the CBKR+ framework to a real world scenario.

In addition to the above, Corpus Based Knowledge Representation is effective for a certain class of applications. The enhancement proposed by this thesis – the CBKR+ framework – extends this class of applications. Additional applications can make use of this approach if it is further enhanced by the SOKR and ISA frameworks explained next:

- The methodology of Self Organizing Knowledge Representation [SOKR] is based on Neural Networks and other such algorithms that have their roots in biological backgrounds. Self-organization is an alternative paradigm to static knowledge representation methods. This paradigm addresses the brittleness and rigidity of representational systems by enabling them to adapt dynamically in order to handle unspecified domains. [Iyengar & Bastani, 1992]
- The Information Systems Architecture [ISA] model is a matrix of five rows and six columns that views a given domain or representational system from many viewpoints. Figure 5.1, redrawn from the source [Figure 6, Pages. 600-601, Sowa & Zachmann, 1992], is a diagrammatic representation of the components of this framework.



Perspective Level	What Entity/ Relationship	How Function/ Argument	Where Location/ Link	Who Agent/ Work	When Time/ Cycle	Why Ends/ Means
Scope						
Enterprise Model						
System Model						
Technology Model						
Component						
Working System	Data	Function	Network	Organization	Schedule	Strategy

Figure 5.1.: Information Systems Architecture

[Figure 6, Pages. 600-601, Sowa & Zachmann, 1992]

The ISA model portrays thirty of these perspectives on knowledge representation, and shows their inter-relation. It works on the assumption that the concrete aspects of the domain must be related to the abstract bits in the computer for useful processes to take place. [Zachmann, 1987]

This research suggests that incorporating Self-Organizing Knowledge Representation [SOKR] paradigms into the CBKR+ framework will help it to dynamically adapt to a changing domain. The axioms of the Information Systems Architecture [ISA] model will provide a concrete structure for the categorization of information in the SOKR methodology. It will also provide a framework of relationships between the various components of a domain, which will again assist the SOKR method by possibly reducing the complexity of implementation and the space explosion drawback.

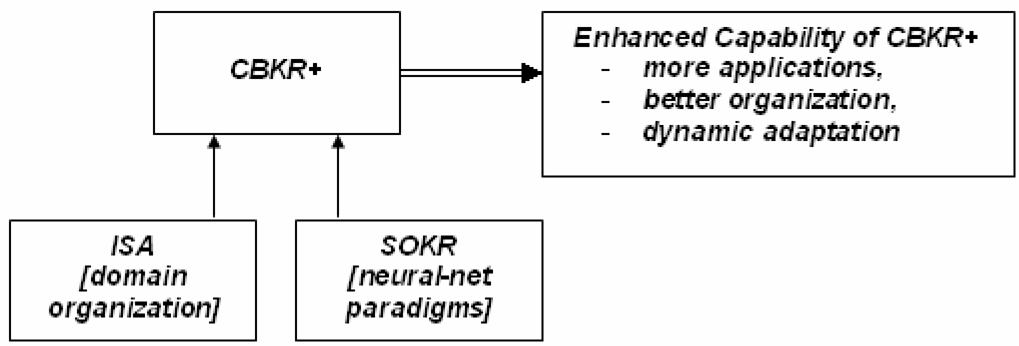


Figure 5.2.: Enhanced CBKR+ using SOKR and ISA

Figure 5.2 above, shows how the three – the ISA model, the SOKR methodology and the CBKR+ framework – may combine to form an efficient, flexible, and widely applicable methodology of knowledge representation. The final result of self organization of the knowledge in the CBKR+ model, assisted by the ISA framework, will increase the types of applications that can be addressed by including those that benefit from a well-defined domain representation, fine-tune the corpus by organizing both components and their inter-relations using neural learning paradigms, and reduce the brittleness of the corpus by giving it the ability to adapt to a changing domain dynamically. The possibility of this union of methodologies to allow the establishment of relations between components of the schemas of different domains can also be explored.

5.4. Contributions of the Thesis

The CBKR methodology used statistics defined for CBKR to address a limited class of applications. The methodology had two identified drawbacks: (1) limited association capability and (2) no support for hypothesis testing and prediction in context. The lack of association capability arose from the fact that the methodology did not have criteria in place to extract substantial associations in the corpus. The lack of support for hypothesis testing and prediction in context limited what could be analyzed using the CBKR approach to query-processing, without support for analysis and inference from the corpus.

Through this work, the researcher provided a solution to the drawbacks identified above in the form of the proposed Corpus Based Knowledge Representation Plus [CBKR+] framework. The framework increased the expressiveness of the corpus-based approach by identifying association criteria, and incorporating them with the CBKR methodology. This was done to allow the support of new forms of analyses related to hypothesis testing and prediction in context that could be applied to domains such as criminology. This framework revealed a variety of properties of the domain, and helped to improve the completeness and expressiveness of the domain's representation. The new framework led to the following contributions of this thesis:

- Improved the expressiveness of the representation by identifying associations in the corpus of schema representations using a suggested Set of Criteria for Association,
- Interpreted a traditional domain categorization model, called Brian Gaines' Domains of Reasoning model [Gaines, 1987], to define a new domain categorization model called the Basis for Categorization model,
- Incorporated the Basis for Categorization model to (1) facilitate a first level categorization of the schema components in the corpus into possible associations and association makers, and (2) define the Set of Criteria for Association, so that the set covered all types of associations and association agents,
- Defined analysis mechanisms to identify and extract further associations in the corpus through the suggested Set of Criteria for Association, so as to perform hypothesis testing and prediction in context,
- Demonstrated the framework by (1) showing how the new approach can enhance the original CBKR methodology through examples from CBKR, and (2) applying the framework to the physical world domain of Criminology.

This work is significant as it offers a simple and structured methodology to organize knowledge in the form of the representational components of a given domain. It makes available an organized framework to create associations within the components of schemas that represent a domain, which can be analyzed and used for various applications. This research provides an approach that is more expressive, and supports deeper analyses through more diagnostic and probability-based forms of queries. It will be of interest to data analysts and information specialists as it offers a simple method to collate and organize knowledge in the form of the schema components of a given domain that can be analyzed, and used for various applications. This work will also be of significance to database developers and soft computing specialists, as it makes available an organized framework to store the schema components of the representations of a domain. Lastly, this research will be of interest to socio-technical research professionals, criminologists and profilers who wish to leverage the automation that computer science can impart to the domain of Criminology in order to reduce the manpower required for the process.

Glossary of Terms

Basis of Categorization model

The Basis of Categorization model is developed in this work as an instance of the Domains of Reasoning model. It creates a first-level categorization of the components of the corpus into possible associations and association makers.

Brittleness

Brittleness [of a knowledge representation] indicates that the representation of a given domain is incomplete due to the inability to model a domain in its entirety as compared to the physical world.

Component

In this work, a component is a single element of the finest granularity of the corpus, where the granularity is decided by the designer of the corpus to suit the domain of application.

Corpus

Each schema that represents a given domain is composed of components that model the domain, and associations between those components. If a library of such schemas is constructed, and mappings, or associations, are established between the various components of the representations, it is called a corpus. [Halevy & Madhavan, 2003]

Corpus Based Knowledge Representation [CBKR]

The Corpus Based Knowledge Representation is a method that builds and uses a library of domain representations and schemas and the knowledge contained within them. [Halevy & Madhavan, 2003]

Corpus Based Knowledge Representation [CBKR] Methodology

CBKR is called a methodology because it is a set of rules and procedures applied to a corpus of schemas that represent a domain.

Corpus Based Knowledge Representation Plus [CBKR+]

As defined by this work, CBKR+ is a basic conceptional structure for classifying and organizing the components in the corpus.

Corpus Based Knowledge Representation Plus [CBKR+] Framework

CBKR+ is called a framework because it incorporates the Basis of Categorization model and the Set of Criteria for Association with CBKR in order to enhance the technique.

Domain

A domain is any particular part of the world that must be represented for computational use. [Halevy & Madhavan, 2003]

Domains of Reasoning model

The Domains of Reasoning model outlines some foundational considerations for the logical organization and representation of knowledge in a given domain system. [Gaines, 1987]

Knowledge representation

A knowledge representation consists of a methodology for computer systems to be able to encode and utilize information about the domain under consideration. [Stone, 2003]

Schema

A schema is a specific way to document a given domain, and model its structure and content. [Doan et al., 2001, Madhavan et al., 2003]

Set of Criteria for Association

The Set of Criteria for Association developed in this work includes eight factors or criteria that can be used to identify and extract associations between the components of the corpus.

Bibliography

[Aked, 2005]

Aked J. Joy Aked's Forensic Psychology Page. Lecture notes for Forensic Psychology – PSY 3013/4013, University of Paisley.

http://www-socsci.paisley.ac.uk/JA/Profiling.doc [Current: August, 2006]

[American Heritage Dictionary, 2000]

The American Heritage® Dictionary of the English Language, Fourth Edition, © 2000, by Houghton Mifflin Company.

[Anderson, 1995]

Anderson J. 1995. Cognitive Psychology and its Implications. New York: W.H. Freeman and Company.

[Androutsopoulos et al., 1995]

Androutsopoulos I., Ritchie G., Thanisch P. 1995. Natural Language Interfaces to Databases: An Introduction. Journal of Language Engineering, Vol. 1, Part 1, Pgs: 29-81.

[Aronson & Rindflesch, 1998]

Aronson A., Rindflesch T. (1998) Semantic Knowledge Representation Project. A Report to the Board of Scientific Counselors, Cognitive Science Branch, Lister Hill National Center for Biomedical Communications.

[Baets, 1998]

Baets W. 1998. Organizational Learning and Knowledge Technologies in a Dynamic Environment. Kluwer Publishers.

[Bernstein, 2003]

Bernstein P. 2003. Applying Model Management to Classical Metadata Problems. Proceedings of the First Biennial Conference on Innovative Data Systems Research.

[Bichindaritz & Conlon, 1996]

Bichindaritz I., Conlon E. 1996. Temporal Knowledge Representation and Organization for Case-Based Reasoning. Proceedings of the Third Workshop on Temporal Representation and Reasoning.

[Brachman & Levesque, 1985]

Brachman R, Levesque H. 1985. Readings in Knowledge Representation. Morgan Kaufman, Los Altos.

[Brewka, 1991]

Brewka G. 1991. Nonmonotonic Reasoning: Logical Foundations of Commonsense. Cambridge University Press.

[Davis, 1991]

Davis R. 1991. A Tale of Two Knowledge Servers. AI Magazine, Vol. 12, Issue 3, Pgs: 118-120.

[Davis & Shrobe, 1983]

Davis R., Shrobe H. 1983. Representing Structure and Behavior of Digital Hardware. IEEE Computer, Special Issue on Knowledge Representation, Vol. 16, Issue 10, Pgs: 75-82.

[Davis et al., 1993]

Davis R., Shrobe H., Szolovits P. 1993. What is a Knowledge Representation? AI Magazine, Vol. 14, Issue 1, Pgs: 17-33.

[Delugach & Rochowiak, 2001]

Delugach H., Rochowiak D. 2001. Lecture Notes for CS630/730: Artificial Intelligence I/II. University of Alabama, Huntsville.

[Doan et al., 2001]

Doan A., Domingos P., Halevy A. 2001. Reconciling Schemas of Disparate Data Sources: A Machine Learning Approach. Proceedings of ACM SIGMOD International Conference on Management of Data, Pgs: 509-520, 2001.

[Doan et al., 2002]

Doan A., Madhavan J., Domingos P., Halevy A. 2002. Learning to Map Between Ontologies on the Semantic Web. Proceedings of the Eleventh International WWW Conference, 2002.

[Easton & McColl, 2005]

Easton V., McColl J. Confidence Intervals and Hypothesis Testing. http://www.cas.lancs.ac.uk/glossary_v1.1/ [Current: August, 2006]

[Evermann & Wand, 2005]

Evermann J., Wand Y. 2005. Toward Formalizing Domain Modeling Semantics in Language Syntax. IEEE Transactions on Software Engineering, Vol. 31, No. 1.

[Faid et al, 1999]

Faid M., Missaoui R., Godin R. 1999. Knowledge Discovery in Complex Objects. Computational Intelligence, Vol. 15, No. 1.

[Gaines, 1987]

Gaines B. 1987. Logical Foundations for Knowledge Representation in Intelligent Systems. Proceedings of the ACM SIGART International Symposium on Methodologies for Intelligent Systems, Pgs: 366–380.

[Ganter & Wille, 1999]

Ganter B., Wille R. 1999. Formal Concept Analysis: Mathematical Foundations. Springer-Verlag.

[Grzymala-Busse, 1986]

Grzymala-Busse J. 1986. Algebraic Properties of Knowledge Representation Systems. Proceedings of the first ACM SIGART International Symposium on Methodologies for Intelligent Systems, Pgs: 432-440.

[Halevy & Madhavan, 2003]

Halevy A., Madhavan J. 2003. Corpus Based Knowledge Representation. Proceedings of the International Joint Conference on Artificial Intelligence, Pgs: 1567-1572.

[Halevy et al., 2003]

Halevy A., Etzioni O., Doan A., Ives Z., Madhavan J., McDowell L., Tatarinov I. 2003. Crossing the Structure Chasm. Proceedings of the First Biennial Conference on Innovative Data Systems Research.

[Hayes, 1985a]

Hayes P. 1985. Some Problems and Non-Problems in Representation Theory. Readings in Knowledge Representation, Pgs: 3-22. Morgan Kaufmann.

[Hayes, 1985b]

Hayes P. 1985. The Logic of Frames. Readings in Knowledge Representation, Pgs: 287-295. Morgan Kaufmann.

[Hoffmann, 1997]

Hoffmann M. 1997. Is there a "Logic" of Abduction? Proceedings of the Sixth Congress of the International Association for Semiotic Studies.

[Iyengar & Bastani, 1992]

Iyengar S., Bastani F. 1992. Self-Organizing Knowledge and Data Representation in a Distributed Environment (Guest Editors' Introduction). IEEE Transactions on Knowledge and Data Engineering, Vol. 4, No. 2.

[Jigzone, 2006]

Jig-Saw Puzzle. http://www.jigzone.com/z.php?i=C4055D50F732&z=6 [Current: August, 2006]

[Kwok et al., 2001]

Kwok C., Etzioni O., Weld D. 2001. Scaling Question Answering to the Web. Proceedings of the Tenth International WWW Conference.

[Lau & Horvitz, 1999]

Lau T., Horvitz E. 1999. Patterns of Search: Analyzing and Modeling Web Query Refinement. Proceedings of the seventh International Conference on User Modeling, Pgs: 119-128.

[Levesque & Lakemeyer, 2000]

Levesque H., Lakemeyer G. 2000. The Logic of Knowledge Bases. MIT Press, Cambridge, Massachusetts.

[Lieberman & Liu, 2002]

Lieberman H., Liu H. 2002. Adaptive linking between Text and Photos using Common Sense Reasoning. Lecture Notes in Computer Science 2347 on Adaptive Hypermedia and Adaptive Web-Based Systems, Pgs: 2-11.

[Madhavan et al., 2001]

Madhavan J., Bernstein P., Rahm E. 2001. Generic Schema Matching with Cupid. Proceedings of the Twenty–seventh VLDB Conference.

[Madhavan et al., 2002]

Madhavan J., Bernstein P., Domingos P., Halevy A. 2002. Representing and Reasoning about Mappings between Domain Models. Eighteenth National Conference on Artificial Intelligence, Pgs: 80-86.

[Madhavan et al., 2003]

Madhavan J., Bernstein P., Chen K., Halevy A., Shenoy P. 2003. Matching Schemas by Learning from Others. Working notes of the IJCAI-03 Workshop on Data Integration on the Web.

[Madhavan et al., 2003a]

Madhavan J., Bernstein P., Chen K., Halevy A., Shenoy P. 2003. Corpus based Schema Matching. Proceedings of IJCAI-03 Workshop on Information Integration on the Web, Pgs: 59-63.

[Markman, 1999]

Markman A. 1999. Knowledge Representation. Mahwah, NJ: L. Erlbaum.

[Merriam-Webster, 2006a]

Merriam-Webster Online Dictionary by Merriam-Webster, Incorporated. http://www.m-w.com/dictionary/methodology [Current: August, 2006]

[Merriam-Webster, 2006b]

Merriam-Webster Online Dictionary by Merriam-Webster, Incorporated. http://www.m-w.com/dictionary/framework [Current: August, 2006]

[Minsky, 1975]

Minsky M. 1975. A Framework for Representing Knowledge. The Psychology of Computer Vision. McGraw-Hill, New York, USA.

[Mohan & Kashyap, 1988]

Mohan L., Kashyap R. 1988. An Object-Oriented Knowledge Representation for Spatial Information. IEEE Transactions on Software Engineering, Vol. 14, No. 5.

[Mylopolous, 1980]

Mylopoulos J. 1980. An Overview of Knowledge Representation. Proceedings of the workshop on Data Abstraction, Databases and Conceptual Modeling, International Conference on Management of Data, Pgs: 5-12.

[Norton, 2005]

Norton H. Types of Reasoning. Lecture Notes for Speech Communication 100C: Message Analysis, Pennsylvania State University.
http://www.personal.psu.edu/faculty/h/m/hmn109/Courses/100C/Reasoning.html [Current: August, 2006]

[Papadimitriou, 1996]

Papadimitriou C. 1996. The Complexity of Knowledge Representation. Proceedings of the Eleventh IEEE Conference on Computational Complexity.

[Partridge, 1996]

Partridge D. 1996. Knowledge Representation. Handbook of Perception and Cognition, Vol. 14, Computational Psychology and AI. Academic Press, NY, Pgs: 55-87.

[Peat & Willet, 1991]

Peat H., Willet P. 1991. The Limitations of Term Co-occurrence Data for Query Expansion in Document Retrieval Systems. Journal of the ASIS, Vol. 42, No. 5, Pgs: 378-383.

[Pesonen, 2002]

Pesonen J. 2002. Concepts and Object-Oriented Knowledge Representation. Master's Thesis. University of Helsinki.

[Petherick, 2005]

Petherick W. Criminal Profiling 1, Criminal Profiling 2, Cyberstalking, Predicting Dangerousness, and Victimology. http://www.crimelibrary.com/about/authors/petherick/index.html [Current: August, 2006]

[Pinker, 1997]

Pinker S. 1997. How the Mind Works. Norton, NY.

[Polanyi, 1983]

Polanyi M. 1983. The Tacit Dimension. Gloucester, Massachusetts.

[Popescu et al., 2003]

Popescu A., Etzioni O., Kautz H. 2003. Towards a Theory of Natural Language Interfaces to Databases. Proceedings of the Conference on Intelligent User Interfaces, 2003.

[Radev et al., 2002]

Radev D., Fan W., Qi H., Wu H., Grewal A. 2002. Probabilistic Question Answering on the Web. Proceedings of the Eleventh International WWW Conference.

[Raghavan & Tsaparas, 2002]

Raghavan P., Tsaparas P. 2002. Mining Significant Associations in Large Scale Text Corpora. Proceedings of the IEEE International Conference on Data Mining, Pgs: 402-409.

[Rasmequan, 2001]

Rasmequan S. 2001. An Approach to Computer-based Knowledge Representation for the Business Environment using Empirical Modeling. Technical Report, Department of Computer Science, University of Warwick.

[Reid, 1961]

Reid L. 1961. Ways of Knowledge and Experience. George Allen & Unwin, London.

[Rieger, 1997]

Rieger B. 1997. Computational Semiotics and Fuzzy Linguistics: On Meaning Constitution and Soft Categories. Proceedings of the International Conference on Intelligent Systems and Semiotics, Pgs: 541-551.

[Rumelhart & Norman, 1988]

Rumelhart D., Norman D. 1988. Representation in Memory. Stevens' Handbook of Experimental Psychology, 2nd Edition, Vol. 2. John Wiley & Sons.

[Russell, 1972]

Russell B. 1972. The Problems of Philosophy. Oxford University Press.

[Schank & Abelson, 1977]

Schank R., Abelson R. 1977. Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures. Lawrence Erlbaum Associates, Hillsdale, NJ.

[Selie, 1996]

Selie P. 1996. The Analysis of Special Psychological Profile: Beginning of Profiling in Slovenia. National Criminal Justice Reference Service's Proceedings on Policing in Central and Eastern Europe: Comparing Firsthand Knowledge with Experience from the West.

[Sizov et al., 2003]

Sizov S., Biwer M., Graupmann J., Siersdorfer S., Theobald M., Weikum G., Zimmer P. 2003. The BINGO! System for Information Portal Generation and

Expert Web Search. Proceedings of the First Biennial Conference on Innovative Data Systems Research.

[Smith, 1982]

Smith B. 1982. Prologue to Reflection and Semantics in a Procedural Language. Readings in Knowledge Representation (Published 1985). Los Altos, CA: Morgan Kauffman 31-39.

[Solan et al., 2005]

Solan Z., Horn D., Ruppin E., Edelman S. 2005. Unsupervised Learning of Natural Languages. Proceedings of the National Academy of Sciences, Vol. 102, No. 33.

[Sowa, 2000]

Sowa J. 2000. Knowledge Representation – Logical, Philosophical, and Computational Foundations. Brooks & Cole, CA.

[Sowa & Zachmann, 1992]

Sowa J., Zachmann J. 1992. Extending and Formalizing the Framework for Information Systems Architecture. IBM Systems Journal Vol. 31, Issue 3, Pgs: 590-616.

[Stark, 2005]

Stark M. Notes on Experimental Statistics: About Statistics. http://www.cs.umd.edu/~mstark/exp101/statistics.html [Current: August, 2006]

[Stone, 2003]

Stone A. 2003. Knowledge Representation for Language Engineering. Handbook for Language Engineers. Center for the Study of Language Information.

[Thagard, 1996]

Thagard P. 1996. Mind: Introduction to Cognitive Science. MIT Press, Cambridge, MA.

[Torsun, 1995]

Torsun I. 1995. Foundations of Intelligent Knowledge-Based Systems. Academic Press, San Diego.

[Utrecht, 2005]

Knowledge Based Systems: Approximate Reasoning. Intelligent Systems Group, Department of Information and Computing Sciences, Utrecht University, Netherlands. http://www.cs.uu.nl/groups/IS/kbs/kbs.html#approx [Current: August, 2006]

[Waltz & Kasif, 1995]

Waltz D., Kasif S. 1995. On Reasoning from Data. ACM Computing Surveys, Vol. 27, No. 3, Pgs: 356-359.

[Way, 1994]

Way E. 1994. Knowledge Representation and Metaphor. Intellect Books.

[Webster’s, 1998]

Webster's Revised Unabridged Dictionary, © 1998 MICRA, Inc.

[Williams & Colomb, 2002]

Williams J., Colomb G. 2002. The Craft of Argument, 2nd Ed. Longman Publishing Group.

[Wirth, 2005]

Wirth U. 2002. What is Abductive Inference? http://www.rz.uni-frankfurt.de/~wirth/inferenc.htm [Current: August, 2006]

[Woods, 1990]

Woods W. 1990. Important Issues in Knowledge Representation. Expert Systems: A Software Methodology for Modern Applications, IEEE Computer Society Press, Pgs: 180-204.

[Yokasu, 2005]

Yosaku. Nishiwaki Lab, Japan Association for Philosophy of Science. http://www.cs.uu.nl/groups/IS/kbs/kbs.html#approx [Current: August, 2006]

[Zachmann, 1987]

Zachmann J. 1987. A Framework for Information Systems Architecture. IBM Systems Journal Vol. 26, Issue 3, Pgs: 276-292.

[Zachman, 1997]

Zachman, J. 1997. Concepts of the Framework for EA - Background, Description and Utility.

http://www.ies.aust.com/~visible/papers/zachman3.htm [Current: August, 2006]