

Using Agents for Unification of Information Extraction and Data Mining

T04320

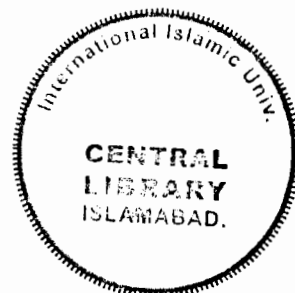


Developed by
Sharjeel Imtiaz
(108-CS/MS/2003)
Azmat Hussain
(124-CS/MS/2003)

Supervised by
Prof. Dr. M. Sikandar H. Khiyal

Department of Computer Science
Faculty of Basic & Applied Sciences

**International Islamic University, Islamabad
(2007)**



~~scribble~~



2
m.p

Accession No TH-2320

MS

006.312

SHU

D.F.
14.12.10

- 1- Database management
- 2- Data mining
- 3- Intelligent agents (computer software)

**Faculty of Basic & Applied Sciences
Department of Computer Science**

Dated: _____

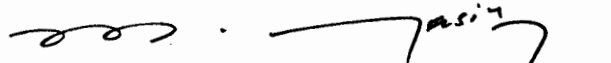
Final Approval

It is certified that we have read the thesis titled "Using Agents for Unification of Information Extraction and Data Mining" submitted by Sharjeel Imtiaz, Reg No. 108-CS/MS/2003 and Azmat Hussain, Reg No. 124-CS/MS/2003. It is our judgment that this thesis is of sufficient standard to warrant its acceptance by the International Islamic University, Islamabad, for the Degree of MS in Computer Science.

Committee

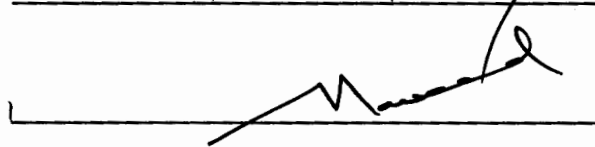
External Examiner

Dr. Mehboob Yasin



Internal Examiner

Dr. Naveed Ikram



Supervisor

Prof. Dr. M.Sikanar H. Khiyal

Chairperson

Department of Computer Science

Fatima Jinnah Women

University, The Mall,

Rawalpindi



بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

In the name of Allah (SWT) the most beneficent and the most merciful.

*A dissertation submitted to the
Department of Computer Science,
International Islamic University, Islamabad
as a partial fulfillment of requirements
for award of the degree of
MS in Computer Science.*

Dedicated...

*To our Parents,
And
To whom we love and respect.*

Declaration

We hereby declare that this software and the thesis, neither as a whole nor as a part thereof, has been copied out from any source. It is further declared that we have developed this software and thesis entirely on the basis of our personal efforts made under the sincere guidance of our teachers and supervisor. No portion of the work presented in this report has been submitted in support of any application for any other degree or qualification of this or any other university or institute of learning.

Sharjeel Imtiaz

108-CS/MS/2003

Azmat Hussain

124-CS/MS/2003

Declaration

We hereby declare that this software and the thesis, neither as a whole nor as a part thereof, has been copied out from any source. It is further declared that we have developed this software and thesis entirely on the basis of our personal efforts made under the sincere guidance of our teachers and supervisor. No portion of the work presented in this report has been submitted in support of any application for any other degree or qualification of this or any other university or institute of learning.

Sharjeel Imtiaz

108-CS/MS/2003

Azmat Hussain

124-CS/MS/2003

Acknowledgement

It is great blessing of Allah Almighty, who gave us courage to accomplish this task. We are grateful to our supervisor **Dr. M. Sikandar H. Khiyal** for his keen dedication to accomplish this work. With the great guidance and self believe on us we accomplish our research according to given direction.

I particularly thank to **Maccallum** who was our main source of inspiration for this idea and motivation.

The great effort we show in **International Conference** which was our main achievement under the guidance of our teacher in 2005 at Karachi.

We are really grateful to our parents who put confidence on us and their patience was admirable to accomplish this research.

Sharjeel Imtiaz
(108-CS/MS/2003)
Azmat Hussain
(124-CS/MS/2003)

Project in Brief

- Project Title:** Using Agents for Unification of Information Extraction and Data Mining
- Objective:** To provide unification of information extraction and data Mining techniques for special domain Cancer, HIV related proteins
- Undertaken By:** Sharjeel Imtiaz
108-CS/MS/2003
Azmat Hussain
124-CS/MS/2003
- Supervised by:** Prof Dr, M.Sikandar H. Khiyal

Chairperson
Department of Computer Science
Fatima Jinnah Women
University , The Mall,
Rawalpindi
- Data Started:** September 01, 2004
- Date Completed:** July, 2007
- Tools Used:** Java, JADE, ORACLE 10g/ Warehouse Builder

Visio 2005, J Builder 2005
- Operating Systems:** Microsoft Window XP/NT/2000

Abstract

Early work for unification of information extraction and data mining is motivational and problem stated work. This research proposes a solution framework for unification using intelligent agents. It also provides strong unification and new K_Mean algorithm for cancer, HIV related and general proteins. The unification technique could unify any data mining technique using feature matrix. The Jaccard approach further measures dissimilarity value for each class of Cancer, HIV and general proteins. The agents communicate horizontally and vertically to complete the unification by collaboration and coordination.

Table of Contents

Ch. No	Contents	Page No
1	Introduction	1
1.1	Problem Statement.....	1
1.2	Purpose of Research.....	2
1.3	Conventional Unified Model	2
1.4	Unification Modeling.....	2
1.5	Unification Framework and Advance Technology.....	3
1.5.1	Jade Tool kit.....	3
1.5.2	Agents and Information Extraction.....	3
1.5.3	Jade and J2EE	3
1.6	Machine Learning Model.....	4
1.6.1	Label Sequential Data.....	4
1.6.2	Undirected Graphical Models	5
1.6.3	Potential Functions.....	5
1.6.4	Conditional Random Fields	6
1.7	Clustering.....	7
1.7.1	Partitioning Methods.....	7
1.7.2	Hierarchical Methods.....	7
1.7.3	Density based Models.....	8
1.7.4	Grid based Methods	8
1.7.5	Model based Methods.....	8
2	Basic Theory	9
2.1	Integration of Information Extraction and Data Mining.....	9
2.2	Unification and Architecture Approach.....	9
2.3	Problem Domain	10
2.4	Functional Requirements	11
2.4.1	Processing of Medline Abstracts	11
2.4.2	Preprocessed Document Views.....	11
2.4.3	Feature Extraction and Annotation	11
2.4.4	Classification Procedure	12
2.4.5	Query Submission (Optional)	12
2.5	Non Functional Requirements	12
2.5.1	User Interface and Human Factors	12
2.5.2	Enhancement.....	13
2.5.3	Interpretability.....	13
2.5.4	Hardware Specification.....	13
2.5.5	Software Specification.....	13
2.5.6	Performance Characteristics	13
2.6	Problem Analysis	13
2.6.1	Approaches	13
2.6.2	Why Warehouse.....	14
2.6.3	Agent Architecture.....	14
2.6.4	Excluded Areas	14
2.6.5	Related Areas	15
2.6.6	Sentence Structure	15
2.6.7	Protein Sentence Structure.....	15
2.7	Comparison with other Techniques	16

2.8	Top Down and Bottom up Approach.....	16
3	Literature Survey	18
3.1	Interactive Information Extraction with Constrained Conditional Random Fields [12]	18
3.2	A Note on the Unification of Information Extraction and Data Mining using Conditional-Probability, Relational Models [1]	18
3.3	A Mutually Beneficial Integration of Data Mining and Information Extraction [8].....	19
3.4	Mining Soft-Matching Rules from Textual Data [3]	19
3.5	An open Architecture [4]	19
3.6	FASTUS: A Cascaded Finite-State Transducer for extracting information from Natural language Text [32]	20
3.7	Shallow Parsing with Conditional Random Fields [10]	20
3.8	Mining Association Rules between Sets of Items in Large Databases [27]	20
3.9	Iterative Classification in Relational Data [28]	21
3.10	Joins that Generalize: Text Classification Using WHIRL [24]	22
3.11	Improving the Representation of Legal Case Texts with Information Extraction Methods [30].....	23
3.12	A Novel Use of Statistical Parsing to Extract Information from Text [29]	23
3.13	Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching [31]	24
3.14	Learning to Combine Trained Distance Metrics for Duplicate Detection in Databases [26].....	25
3.15	Automating the Construction of Internet Portals with Machine Learning [25].	25
3.16	Automatic Extraction of Protein Interactions from Scientific Abstracts [17] ...	26
3.17	A Web-resource for Exploring Protein Co-occurrences in MEDLINE Abstracts [18] 27	
3.18	Unsupervised Gene/Protein named Entity Normalization using Automatically Extracted Dictionaries [19]	28
3.19	A Protein Interaction Extraction System [20].....	28
3.20	Automatic Extraction of Gene and Protein Synonyms from MEDLINE and Journal Articles [21]	29
3.21	Comparison Summary	30
4	Methodology.....	31
4.1	Prior Approaches	31
4.1.1	Machine Learning technique.....	31
4.1.2	Probabilistic Models	32
4.1.3	Generative Models	32
4.1.4	Prior Agent Model	32
4.1.5	K-Mean Algorithm.....	33
4.2	Our Approaches	34
4.2.1	Agent Architecture and Messaging.....	34
4.2.2	Detail Architecture	35
4.2.3	Unification Procedure	37
4.2.4	Hybrid Approach of K-mean and Jaccard Approach.....	37
4.2.5	Jaccard Approach.....	38
4.2.6	New K-Mean Algorithm	39
4.3	Cyclicity of Data	41
4.4	Infrastructure of Application.....	42
4.5	Algorithm/ Pseudo Code.....	43

5	Implementation / Results	49
5.1	Application.....	49
5.1.1	Star Schema	49
5.1.2	Transformation.....	51
5.1.3	Volume of Data.....	51
5.1.4	Physical Issues	52
5.1.5	Profiling	53
5.1.6	Operating System Issues	53
5.1.7	Nature of Ware House	53
5.1.8	Class Diagram	54
5.1.9	Jaccard Approach Implementation	59
5.1.10	Features Agent Implementation.....	59
5.1.11	Rule Checking by Feature Agent Ancestor	59
5.1.12	Special Method and Feature Building.....	59
5.2	Collaboration Diagrams	59
5.2.1	Black Box to Feature Agent.....	60
5.2.2	Feature Agent to Morphological Agent	60
5.2.3	Morphological Agent to Feature Agent	61
5.2.4	Feature Agent to Pos tag Agent	61
5.2.5	Pos tag Agent to Feature Agent	61
5.2.6	Feature Agent to black box	62
5.2.7	Monitor Agent to Data Modeler Agent.....	62
5.2.8	Data Agent to Data Modeler Agent	62
5.2.9	Data Modeler Agent to Black Box Agent.....	63
5.2.10	Monitor Agent to Data Modeler Agent.....	63
5.2.11	Data Modeler Agent to Data Agent	63
5.2.12	Data Modeler Agent to Text Agent	64
5.3	OS (Operating System) Settings	64
5.4	Batch Processes.....	64
5.5	Error Handling	65
5.6	Agent Communication	66
5.7	Results.....	67
5.7.1	Medline Abstracts Main Screen.....	69
5.7.2	Black Box Agent and Classification	70
5.7.3	Classification Results.....	71
5.7.4	Classification Statistics	72
5.7.5	Analysis of Results	73
5.8	Conclusion	79
5.9	Future Direction	80
Appendix A Implementation Code		81
1.	Black Box Agent and Classification	81
2.	Jaccard Approach Implementation	82
3.	Features Agent Implementation.....	83
4.	Rule Checking by Feature Agent Ancestor	84
5.	Special Method and Feature Building.....	85
Appendix B K-Mean Analysis using SPSS		87
Appendix C Abner Constraint Random Field Tool (CRF)		88
Bibliography & References		89

Chapter 1

Introduction

1 Introduction

Unified model is based on undirected graphical model; the current model is motivational work. Primary outcome of unification is common inference procedure [1]. Approaches like; soft matching, data mining and unification models developed but unification is weak [2,3]. Weak unification means that unification among two domains by more than two models.

1.1 Problem Statement

The research of Maccallum is motivational work for Unification. Unification has four common properties [1]:

- Common Procedure
- Common Outcome
- Common Output
- Common Input

Conventional data mining technique has problems like uncertain about extraction of feature so result is invalid structure. Similarly, information extraction model HMM is limited in granularity of features and in state sequence. The Maccallum Model is computational complex so there is need to measure feature set separately for each problem [1]. Hence, it is N square problem. Few techniques cannot resolve complex domain extraction and concept which is dictionary based and grammatical based techniques because of their limitation of extraction [20,22]. So, there is need of technique which can extract data unlimited features and complex names by common unification model for rich database.

The problem domain is about biological names which consist of complex names to extract. Some of proteins names' identification of prefix and suffix features are fixed. Here CDC2, CYCLIN B1 are proteins complex names [22]. The features about complex names need to be identified. Proteins have other concept like synonyms and their association so there is need to identify their feature [18, 19, 20, 21, 22].

- Features identification of protein names

- Feature identification of synonyms and associations

Software agents are also use for extraction of concept from corpus [3]. The communication should be hierarchal and parallel.

The problem statement stated is to extract the important/complex concept from corpus using unification new model by support of software agents.

1.2 Purpose of Research

The research is about framework of information extraction and data mining. The framework supports tight integration among information extraction (IE) and data mining. The objectives of information extraction and data mining unification are as following.

- Efficient precession and recall.
- The model should be efficient that there is need to measure feature set once but not separately for each type of problem like in Constraint Random Model.
- Extraction of complex names and sentences.
- Utilization of software agents
- A tight integration of two domain information extraction and data mining means the model should fulfill Unification properties.

1.3 Conventional Unified Model

Conventional approaches of unified models of machine learning i.e. Support Vector Machine (SVM) but lacks better recall and precision. Various techniques like boosted wrapper, support vector machine (SVM) are classification techniques. But these techniques lack concept of name extraction. Motivational technique of unification is formed by two domains text mining and data mining [1].

1.4 Unification Modeling

It is Unified Model have two domains information extraction and data mining. Data mining relies on data set and extraction from databases. Information extraction is a machine learning technique to extract names and concepts.

Unification is the technique to extract names and concepts. The sentences are constituted of fields [1]. In fact fields are combined to form complete entity. Unification model has following properties,

- Common input
- Common outcomes
- Common procedure
- Common output

1.5 Unification Framework and Advance Technology

There is need of framework for the information extraction and data mining unification. Agents conventionally use for information extraction and data mining separately. Software agents are use for coupling, support, monitoring and analysis of ontologies. There is need of complete unification framework for the data mining and information extraction.

1.5.1 Jade Tool kit

A fundamental characteristic of multi-agent system is the agents communicate and interaction. Agent characteristics are shown by the exchange of messages. It is crucial that agents should agree on the format and semantics of these messages. Jade follows FIPA standards hence Jade agents can interact with agents written in other languages and running on other platforms.

1.5.2 Agents and Information Extraction

The approach of multi-agent is for information extraction. It solves the problem of information extraction using hierarchical and parallel communication technique. Every parent has predecessors to perform particular task of parallel communication [3].

1.5.3 Jade and J2EE

A GUI runs on own thread (the event-dispatching thread), which allows it to handle and react promptly to events. The user interacts with the GUI with component, button or window. An agent program runs on its own thread, which handles behavior. But it is not

efficient to let other thread to call directly. JADE has provides an appropriate mechanism to manage interactions between the two threads when integrating with the GUI.

1.6 Machine Learning Model

There are following models which are related to extraction.

1.6.1 Label Sequential Data

The concept of assigning label sequence to set of observation sequence is found in many fields; the Bioinformatics, computational linguistics and speech recognition. The task of natural language processing is labeling words in a sentence correspond to part-of-speech (POS) tagging. Each word is labeled with a tag indicating its particular part of speech, found in annotated text e.g. (1) He reckons the current account deficit will narrow to only 1.8 billion in September. The tagging of same sentence is as follows: [PRP He] [VBZ reckons] [DT the] [JJ current] [NN account] [NN deficit] [MD will] [VB narrow] [TO to] [RB only] [# #] [CD 1.8] [CD billion] [IN in] [NNP September].

In this way the labeling of sentence is useful preprocessing step for natural language. POS tags are not only support the words but explicitly indicate structure inherent in language.

The most common method for labeling and segmenting task is the hidden Markov model (HMM) or probabilistic finite-state automata. The purpose of HMM is to identify the most likely sequence of labels in a given sentence. HMM is a generative model that defines a joint probability distribution $p(X, Y)$. The X and Y are random variables respectively ranging over observation sequence and corresponding label sequence.

To define joint distributions over generative model enumerate all possible observation sequences. The most domains are intractable unless observation elements are represented as isolated units and independent from the other elements in an observation sequence. More precisely, the observation element at any given instance directly depends on the state and label at that time. This is an appropriate assumption for few simple data sets however most real-world observation sequences are best representation in terms of multiple interacting features and long-range dependencies.

The representation issue is one of the fundamental problems on labeling sequential data.

Clearly, a model should support tractable inference necessarily; however a model that represents the data without making unwarranted independence assumptions is also desirable. One way of satisfying both these criteria is to use a model that defines a conditional probability $p(Y | x)$ over label sequences given a particular observation sequence x , rather than a joint distribution over both label and observation sequences.

Conditional models are used to label a novel observation sequence x by selecting the label sequence y , which maximizes the conditional probability $p(y|x)$. The conditional nature of such models means that no effort is wasted on modeling the observations. One is free from having to make unwarranted independence assumptions about these sequences; arbitrary attributes of the observation data may be captured by the model, without the modeler having to worry about how these attributes are related.

Conditional random fields (CRFs) are a probabilistic framework for labeling and segmenting sequential data, based on the conditional approach described in the previous paragraph. A CRF is a form of undirected graphical model that defines a single log-linear distribution over label sequences given a particular observation sequence. The primary advantage of CRFs over hidden Markov models is their conditional nature, resulting in the relaxation of the independence assumptions required by HMMs in order to ensure tractable inference.

1.6.2 Undirected Graphical Models

A conditional random field may be viewed as an undirected graphical model, or Markov random field, globally conditioned on X , the random variable representing observation sequences. Formally, we define $G = (V, E)$ to be an undirected graph such that there is a node V corresponding to each of the random variables representing an element Y_i of Y . If each random variable Y_i obeys the Markov property with respect to G , then (Y, X) is a conditional random field. In theory the structure of graph G may be arbitrary, provided it represents the conditional independencies in the label sequences being modeled.

1.6.3 Potential Functions

The graphical structure of a conditional random field may be used to factorize the joint distribution over elements Y vector of Y_i into a normalized product of strictly positive, real-valued potential functions, derived from the notion of conditional independence.

Each potential function operates on a subset of the random variables represented by vertices in G . According to the definition of conditional independence for undirected graphical models, the absence of an edge between two vertices in G implies that the random variables represented by these vertices are conditionally independent given all other random variables in the model.

The potential functions must therefore ensure that it is possible to factorize the joint probability such that conditionally independent random variables do not appear in the same potential function. The easiest way to fulfill this requirement is to require each potential function to operate on a set of random variables whose corresponding vertices form a maximal clique within G . This ensures that no potential function refers to any pair of random variables whose vertices are not directly connected and, if two vertices appear together in a clique this relationship is made explicit. In the case of a chain-structured CRF, such as that depicted in Figure 1, each potential function will operate on pairs of adjacent label variables Y_i and Y_{i+1} .

It is worth noting that an isolated potential function does not have a direct probabilistic interpretation, but instead represents constraints on the configurations of the random variables on which the function is defined. This in turn affects the probability of global configurations – a global configuration with a high probability is likely to have satisfied more of these constraints than a global configuration with a low probability.

The product of a set of strictly positive, real-valued functions is not guaranteed to satisfy the axioms of probability. A normalization factor is therefore introduced to ensure that the product of potential functions is a valid probability distribution over the random variables represented by vertices in G .

1.6.4 Conditional Random Fields

Define the probability of a particular label sequence y given observation sequence x to be a normalized product of potential functions, each of the form

$$\phi_i(x_i, y_i, y_{i+1}, x_{i+1})$$

This is a transition feature function of the entire observation sequence and the labels at positions i and $i-1$ in the label sequence. It is a state feature function of the label at position i and the observation sequence; and j and μ_k are parameters to be estimated from training data. When defining feature functions, we construct a set of real-valued features $B(X, i)$ of the observation to express some characteristic of the empirical distribution of the training data that should also hold of the Model distribution. An example of such a feature is $B(X, i) = (1 \text{ if the observation at position } i \text{ is the word "September" otherwise.}$ Each feature function takes on the value of one of these real-valued observation features $B(X, i)$ if the current state (in the case of a state function) or previous and current states (in the case of a transition function) take on particular values. All feature functions are therefore real-valued.

1.7 Clustering

In general, major clustering methods can be classified into the following categories.

1.7.1 Partitioning Methods

Given a database of n objects or data tuples a partitioning method constructs k partitions of data, where each partition represents a cluster and K less than and equal to n . That is, it classifies the data into k groups, which together satisfy the following requirements: (1) each group must contain at least one object and (2) each object must belong to exactly one group. Notice that the second requirement can be relaxed in some fuzzy partitioning techniques.

Given K group, the number of partitions to construct, a partitioning method creates an initial partitioning. It then uses iterative relocation techniques that attempt to improve the partitioning by moving objects from one group to another. The general criterion of a good partitioning is that objects in the clusters are far apart or very different. There are several of kinds of criteria for judging the quality of partition [25].

1.7.2 Hierarchical Methods

A hierarchical method creates a hierarchical decomposition of given set of data objects. A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed. The agglomerative approach is also called

the bottom–up approach, starts with each object forming a separate group. It successively merges the objects or groups close to one to one another, until all of groups are merged into one or until a termination condition holds. The divisive approach also called the top-down approach starts with all the objects in the same cluster. The every successive iteration on a cluster can be split up into smaller cluster until each object is in one cluster or a termination condition fire.

1.7.3 Density based Models

Most partitioning methods cluster objects based on the distance between objects. Such methods can find only spherical – shaped clusters and encounter difficulty at discovering clusters of arbitrary shapes. Other clustering methods have been developed on the notion of density. Their general idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold; this is for each data point within the given radius has to at least a minimum number of points. Such a method can be used to filter out noise and discover clusters of arbitrary shape [25].

1.7.4 Grid based Methods

Grid based methods quantize the object space into a finite number of cells that form a given structure. All of the clustering operations are performed on the grid structure. The main advantage of this approach is its Fast processing time, which is typically independence of number of data objects and dependent only on the number of cells in each dimension in the quantized space [25].

1.7.5 Model based Methods

Model-based methods hypothesize a model for each of the clusters and find the best fit of the data to the given model. A model-based algorithm may locate clusters by constructing a density function that reflects the spatial distribution of the data points. It also leads to a way of automatically determining the number of clusters based on standard statistics, taking “noise” or outliers into account and thus yielding clustering methods [25].

Chapter 2

Basic Theory

2 Basic Theory

The main purpose of this application is to unify the data mining and information extraction. This architecture is developed specifically for protein extraction system. Many proteins belong to particular specie. Doctor can take advantage from this system by submitting query for protein and its interactions about particular cancer.

2.1 Integration of Information Extraction and Data Mining

Techniques use for data mining but these techniques can be used for unification. The concept of extraction of useful patterns from text is called Name Entity Recognition (NER).

Information extraction (IE) is used for data classification and clustering. In 1990, the purpose of information extraction was to reduce the amount of work on following tasks; Email classification, document classifications and news paper classification.

The survey of journals shows the lacking of extraction of cross reference from copra. Infect statistical models are based upon probability theory [1] [7] [9]. Unification of information extraction and data mining is also present inside [3][5]. Information extraction lacks constraints, co-reference resolution and field extraction. Similarly, data mining hide features so covering uncertainty, limited in feature extraction and mine invalid structure. Unified framework is primarily illustrated for producing multitudes of labels, sequence of labels, binary resolution and cross references [1].

Model for rule base learning and soft matching algorithm is for cross reference, which is base on distance editing [3]. Perhaps, these approaches are used combine or standalone but not as unified common approach.

2.2 Unification and Architecture Approach

Our approach proposed tight integration for the boundaries between IE and data mining. Proposed unified system can be understood as single, large conditionally-trained undirected graphical model. The strict causality among events is not necessarily apparent on set of circumstances appearing between low-level text data and higher-level relational data mining.

Third aspect of architecture is about technology framework, which is agent based communication hierarchical and parallel model. Open agent architecture, which is regarding black box oriented framework. The individual software client agent is to communicate by means of goals posted by black board agent [4].

The high level design and detail design of architecture relies on some of aspects. This research utilizes the k-mean algorithm of clustering in the agent based framework. The approach is completely relies on statistical model integration. The tight integration is completely relies on feature matrix instead of feature measuring by partial probability model Constraint Random Model (CRF).

2.3 Problem Domain

In a hospital a patient admitted after measuring its disease concerning with new patterns of amino acid so form a new protein name. Doctor knows its name but not knows protein completely. Doctor consulted with Medline dictionaries from data warehouse. Doctor found few results but results are incomplete and are not up to dated. Different dimension of data is not extracted and classification could not be possible for the same type of data. Doctor wants to obtain latest data about different dimension. System posses such capabilities which serve the expert/doctor to analyze latest data gather with older data about proteins and its concept regarding diseases.

The latest result could be found from Medline sites but they can't provide the complete dimensions of data.

Following are the problems that found after analyze medline abstract.

- Unification tight integration model
- Some of proteins names identification based on prefix and suffix features are fixed
- Data source is limited to one domain either from corpus or from database
- Word of protein does not contain plural characters
- Cross-reference and cross-resolution
- Specified Length of characters is important for processing or extraction
- Special relationships among proteins
- Context specific relationship words "made" , "contain" , "consist of",

combination” etc identification difficult

- Alphanumeric characters mainly impede during features selections

2.4 Functional Requirements

Following function will be performed by system. The requirement is for high level and sub level functions. Use case will further illustrate technical specification in detail.

2.4.1 Processing of Medline Abstracts

The requirement criticality is high and processing of Medline abstracts is based on following sub function.

- Load Medline Abstracts
- Unload Medline Abstracts

The document will be loaded from text file. The document consists of Medline abstracts. The Medline abstracts have proteins and its interactions about cancer and all other diseases. Next tokenization process of documents paragraph by sentences will be initiated. The process will continue until document title and type of disease identified. In tokenization process complete document will be loaded with all sentences. The unload process will remove current batch from system.

2.4.2 Preprocessed Document Views

User can view document before and after extraction. The document will load from permanent storage.

- Load Stored Structure.

The document will load from permanent storage and in the form of object.

2.4.3 Feature Extraction and Annotation

User will select feature two types morphological and pos tags. The selection of tags will provide particular type of final extraction. The morphological features will extract result about protein. The pos type feature selection will be in the form of annotated text.

- Select Feature Types
- Annotate Text

2.4.4 Classification Procedure

The classification procedure will produce output in the form of clusters and their clusters will be formed for different segments. The cluster will show different type of data. The types are cancer specific proteins, general disease proteins and miscellaneous.

- K-Mean classification results
- Remote Query (Optional)

Remote agent on selection can send query about protein results. The request will be received by remote agent called monitor agent. User will select remote agent option to enable communication among local and remote agent. The remote agent call modeler agent that will initiate the query request but only if remote agent is not selected. The query will be generated for local database. In next release of this research other algorithm will be supported. The store function will permanently store the extracted result or cluster data in database.

2.4.5 Query Submission (Optional)

Query submission will be initiated if user selects remote agent option. The query regarding patient will be submitted to remote or local agent called monitoring agent by doctor/expert.

- Patient Query
- Show Results

The query submitted and result in the form of classified clusters will be output. The show result function will show the output in grid.

2.5 Non Functional Requirements

Non functional requirements of the system are described below.

2.5.1 User Interface and Human Factors

- Special and expert users will use the system for patterns findings.
- All inputs are composed of controls no special bioinformatics device required.

- System will provide stats in the form of graphical form like pie-chart.

2.5.2 Enhancement

The system has capability to enhance from many angles. Similarly feature building and feature matrix formation is input for unification process. Therefore any data mining technique could be integrates with this technique.

2.5.3 Interpretability

The system could be run under any platform without changing implementation details. The directory structure of Linux and Window could be provided without any user input. Similarly, SQL queries are written in case sensitive language to run on Linux.

2.5.4 Hardware Specification

System will run on these requirements

- P-III minimum and recommended P-IV Intel with L3 / L2 cache.
- 128 MB minimum of memory and 256 MB recommended

2.5.5 Software Specification

Java, JADE Tool Kit, Borland JBuilder 6, Oracle 10 g

2.5.6 Performance Characteristics

The complexity of system will measure relatively better than conventionally mean algorithm. Testing techniques measure recall and precession. The k-mean complexity is relatively slower because iterations to minimize square root error mean.

2.6 Problem Analysis

There are following areas which are analyzed for our research particular domain.

2.6.1 Approaches

There are following approaches which are as follows:

- Bayesian directed graph technique
- Soft Mining for Information extraction using Statistical Models
- Constrain Random Field (CRF)
- Rapiere feature based technique
- Harlove Markov Model directed graph technique

- K-mean algorithm for classification
- Support Vector Machine (SVM) for classification

2.6.2 Why Warehouse

This decision support application and management system can be useful for corpus data extraction. This database has star model for protein and its interaction. Star model is variation of snowflake and better than snowflake. The start model has interactions and interactee in the form of proteins. The dimensions types are disease and its types. The patient data is another dimension which is target in start model. The classification summary gets from our new integration algorithm new k-mean. The classification summary will store into other dimension.

2.6.3 Agent Architecture

The architecture comprises of hierarchical and communication approach. The application is base on parallel communication and vertical communication. Architecture covers many of related problems of Natural language processing (NLP). The process flow is starting from central black board agent. All types of interactions will be initiated from black board.

Guide type: Query other agent.

Assists type: Passes requests to all children or to a child.

Volatile type: Save request and response for particular type of agent.

Below is the mapping of all agents to their role type.

Role	Agents
Guide Type	Monitoring agent
Assist Type	Feature agent, Pos agent, Morphological agent, Modeler agent, data agent, text mining agent
Volatile Type	Black box

Table 2.6.3: Illustrate the mappings of agent

2.6.4 Excluded Areas

The areas which are not part of our scope are as follows.

- Genetic algorithm

- Neural network Support Vector Machine (SVM)
- Machine learning rule learning algorithms
- Agent based enforcement learning
- Context free Grammars
- Turing machines
- Parse trees

2.6.5 Related Areas

The approaches and areas on which this research relied are as follows.

- Undirected Graphs
- Partial and conditional Probability
- Information extraction techniques
- Multi agents
- Data mining k – mean algorithm
- Bioinformatics protein Medline abstracts and species analysis

Approaches like Constraint Random field (CRF) [1], Harlove Markov Model (HMM) [17], Conditional Probability based approaches [1, 9, 12, 14], multi agent [4] also used for information extraction for Pos type corpus. Data mining approaches use for extraction of patterns from corpus [3].

2.6.6 Sentence Structure

A sentence structure is composed of complex, compound and complex compound sentences. Sentences compose of grammar and extract by Part of Speech Tag (POS). The pronouns refer to the interaction and to the noun. Complex and compound sentences cannot easily analyze. There are few heuristics to identify pronouns. The noun is important for interaction purpose.

2.6.7 Protein Sentence Structure

Medline abstracts have any type of disease's proteins [22]. e.g. "Anti-proliferate effect of estrogen in breast cancer cells that re-express ER is mediated by aberrant regulation of cell cycle genes". Sentence structure could be of any type subject or object. The sentence contains cell, gene and protein noun. e.g. "These include genes encoding the anti-apoptosis factor *SURVIVIN*, positive cell cycle regulators (*CDC2*, *CYCLIN B1*, *CYCLIN*

B2, CYCLIN G1, CHK1, BUB3, STK6, SKB1, CSE1 L) and chromosome replication proteins (*MCM2, MCM3, FEN1, RRM2, TOP2A, RFC1*).” Here CDC2, CYCLIN B1 and all others are valid protein names. Protein sentence structure contains terms which are interactors, interactee, synonym and all valid related concept.

Protein interaction will be extracted from sentences after analyzing special relationship verbs. Those are “regulate, regulate of, regulate in or to-regulate, regulate by, regulate of in, to regulate of, of to-regulate, to-regulate in, of regulate of, of regulate, by regulate, regulate through, of to- regulate of, regulate of, regulate to, regulate to, regulate of, of regulate in, regulate during”.

Protein and its interaction will also extract information based on linguistic limitations. e.g. “raw text is MLK2 has a role in vesicle formation and its annotation based on MLK2/NN has/VBZ a/DT role/NN in/IN vesicle/NN formation/NN”. It is clearly mentioned that the tag on text are for part of speech reorganization. The linguistic significant combinations can be extracted from parse tree based techniques i.e. S-V-O, S-V, S-V-M and S-M. The combination can be extracted by the link grammar e.g. “The dog chased a cat”. The dog is part of connect phrase and part as connect pairs.

Pronoun feature identifies protein at start of next sentence. The pronoun solves the problem of co-reference or synonyms. e.g. “The SAC6 gene was found by suppression of a yeast actins mutation. Its protein product, Sac6p (previously referred to as ABP67), was independently isolated by actins-filament chromatography and co localizes with actins in vivo”. In above example protein pronoun cross refer to previous sentence by “its” phrase.

2.7 Comparison with other Techniques

The summary of comparison among different technique is in chapter 3. This chapter explains the comparison among protein extraction techniques and unification approach. The extraction mechanism is following patterns and features, which are reference from grammar. But features lack multiplicity, backward and forward looping.

2.8 Top Down and Bottom up Approach

The both approaches have common procedure, common input, common outcome and common output. Agents communicate and collaborate to fulfill the purpose of unification

using matrix. Infect matrix aid model to use top down or bottom up approach, the feature will remain same as common outcome.

Chapter 3

Literature Survey

3 Literature Survey

This chapter contains review of different research papers. These research papers provide us constructive guidelines in this research.

3.1 Interactive Information Extraction with Constrained Conditional Random Fields [12]

Andrew McCallum and David Jensen proposed that Interactive information extraction must cover two aspects. First one is about how to extract field after verification. Secondly, rapid correction of incorrect fields is performed. Regular expression cannot estimate confidence.

3.2 A Note on the Unification of Information Extraction and Data Mining using Conditional-Probability, Relational Models [1]

Razvan Bunescu, Edward M. Marcotte stated that it is a unified frame work for information extraction and data mining. Because information extraction lacks constraints, co-reference resolution and field extraction and it is limited with noisy data. Data mining lacks features for covering uncertainty, limited in feature extraction and invalid structure. Unified framework proposes a common model to solve the problems; multitudes of labels, sequence of labels, binary resolution and cross references.

It is unified framework which extracts data either from Text or from database. It does not care about two different domains rather it has such properties common input, common output and common procedure. There is need of separate model for every type of problem like factorial, affinity, linear model. The factorial model dimension of label measures is like its classes and part of speech (POS). Our approach measures phrase boundaries and part of speech (POS) in same procedure to extract complex names. Hence, our approach supports common point all types of features about part of speech (POS), cross reference and phrase boundary.

3.3 A Mutually Beneficial Integration of Data Mining and Information Extraction [8]

This paper is about integration of KDD technique with IE. This approach of Fei Sha and Fernando Pereira is used for information extraction by using rule learning. It does not handle noise but learn rules to train data and disjoint set of labels validation. Algorithm does not improve the recall and precision but it is for rules so result is either fail or successful.

This algorithm "Rule Mining Phase" is for KDD technique. When extraction of database on rule applied than KDD process initiates to discover patterns. As, it is clearly seen it lacks the unification properties common procedure, common output property.

3.4 Mining Soft-Matching Rules from Textual Data [3]

This research by Un Yong Nahm and Raymond J. Mooney developed an algorithm that will use different types of approaches which is hard mining rules to produce better results. It proposes edit distance learning, similarity metrics and cosine based method for long string detection. It is applied in the form of n-gram indexing and could optimize for short strings. It is motivation to minimize typography errors, misspellings in dirty data.

This algorithm learns rule and new similarity value that will be compared on each extraction if it is better than previous extraction than new rule will be induced. Best rule will be learned but it will be more erroneous where features are specific not generalized enough.

3.5 An open Architecture [4]

Philip R. Cohen, Adam Cheyer define concept of interoperability of system and its subsystem through agent. It defines delegation, data-directed execution, reasoning, planning concepts with traditional architectures like hierarchal execution. This research paper compares traditional architectures with other architectures. This architecture has capability to predict features behavior, reviewing history and rule specification. A temporal language expression evaluated over database.

This architecture is a motivation for our approach as it defines vertical and hierarchical extraction mechanism. But it lacks black board concept so we add black board concept for central controlling mechanism for all types of hierarchal and vertical executions.

3.6 FASTUS: A Cascaded Finite-State Transducer for extracting information from Natural language Text [32]

FASTUS is a set of cascaded finite-state automata developed by Jerry R. Hobbs, Douglas Appelt . FASTUS works at five type's labels complex word, basic phrases, complex phrases, domain events and merging structure. First four apply on sentence by sentence. Hence processing response is reasonable. FASTUS is more effective in case of POS but for other dimensions of label this machine can't predict proteins morphologies.

3.7 Shallow Parsing with Conditional Random Fields [10]

It produces back-forward algorithm like chunker algorithm. Main objective of B. Taskar, P. Abbeel, and D. Koller's algorithm is to find maximum possible features. Shallow parsing applies on second order CRF. It is also named as Gradient conjugates (GC) and voted preceptor produces million of features. However, expectation is on sequence labeling local feature. It estimates Gradient Conjugate expectation of square of global feature by summing square of local features. GC and GIS are used to train CRF.

This model is variation of constraint random field (CRF) and infect it produces better results. Similarly for each dimension of label, there is need of additional turn over for dimensions like phrase boundary, part of speech.

3.8 Mining Association Rules between Sets of Items in Large Databases [27]

Each transaction of customer database consists of items purchased by a customer in a visit. Rakesh Agrawal Tomasz Imielinski_ Arun Swami present an efficient algorithm that generates all significant association rules between items in the database. The algorithm incorporates buyer management and novel estimation and pruning techniques. Rakesh Agrawal Tomasz Imielinski_ Arun Swami also present results of applying this algorithm to sales data obtained from a large retailing company, which shows the effectiveness of the algorithm.

The items set will be analyzed for each transaction T_n . But this item set is belong to database and pre-determined. There is no such rule which determines discrete corpus Text.

Associations measures over frontier test item set on each transaction but it will also contains incorrect data Meta data is not available at the time of prediction. The algorithm will determined the data from predetermined frontier item set so called total precision will be low compared to recall. There is need of strong algorithm which works on both domains database and text.

Our approach is more efficient and works for both domains like text and database. The new approach will predict item set on the basis of feature rather then frontier set which is infect predetermined item sets. Our approach will determine item set on the basis of feature matrix and similarity value will predict data item set.

3.9 Iterative Classification in Relational Data [28]

Relational data offers unique opportunity for improving the classification accuracy of statistical models. If two objects are related, inferring something about one objects one can aid inferencing. We present an iterative classification procedure; it will expose characteristics of relational data. This approach uses simple bayesian classifiers in an iterative fashion and dynamically updating the attributes of objects. Inferences made with high confidence in initial iterations will be feed back into the data.

It measures the performance of approach on binary classification. Experiments indicate that iterative classification significantly increases accuracy when compared to a single-iteration approach.

This research does iterative classification task based on dynamic properties using Bayesian model. Bayesian model is a directed version and we could not feed back to previous label based on dynamic properties. Feed back capabilities are characteristics of constraint random field (CRF), which aid feed back and feed forward capabilities for users. Butthis model [1] is more efficient in measuring of multitude of labels.

Our approach is more efficient to predict dynamic properties from feature matrix for relational data. Multitude of labels is like cross reference (synonyms), phrasal boundaries and part of speech (POS). Hence, Bayesian is a model which is effective where data known prior to prediction and posterior probabilities are unknown.

3.10 Joins that Generalize: Text Classification Using WHIRL [24]

WHIRL is an extension of relational databases that can perform “soft joins” based on the similarity of textual identifiers; these soft joins extend the traditional operation of joining tables based on the equivalence of atomic values. William W Cohen Haym Hirsh evaluates WHIRL on a number of inductive classification tasks using data from the World Wide Web. We show that although WHIRL is designed for more general similarity based reasoning tasks, it is competitive with mature inductive classification systems on these classification tasks. In particular, WHIRL generally achieves lower generalization error than C4.5, RIPPER, and several nearest-neighbor methods. WHIRL is also fast—up to 500 times faster than C4.5 on some benchmark problems. We also show that WHIRL can be efficiently used to select from a large pool of unlabeled items those that can be classified correctly with high confidence.

In this paper some of earlier approaches already have been discussed such as using Rappier by (Cohen 1996). This approach measures the items by vector space similarity measures. Yang and Chute's (1994) measures similarity base on maximum score that is distance weight method. But finally they choose most frequent based method for experiments. But precision lacks in all of above cases. Hence, finally method of score base gets the tuple which have more score and prune later which have fewer score.

This method is base on relation not on features of labels. The data needs some form of relational space to predict further. So it is good for data mining purpose but not for text corpus. Where data is dissimilar and spread over many web sites. Our approach defines common point for both domains. The above approach is also for cross reference problem but not for association/relationship problem.

3.11 Improving the Representation of Legal Case Texts with Information Extraction Methods [30]

The prohibitive cost of assigning indices to textual cases is a major hurdle. While considerable progress has been made toward extracting facts from well-structured text cases or classifying case abstracts.

In this paper, Stefanie Brünninghaus and Kevin D. Ashley show how a representation can facilitate classification-based indexing. They present the hypothesis (1) abstracting from the individual actors and events (2) capturing actions in multi-word features (3) recognizing negation, can lead to a better representation of legal case texts. They discuss how a state-of-the-art information extraction program can be used to implement these techniques. Preliminary experimental results show that a combination of domain-specific knowledge and IE techniques can be used to generalize.

There are three approaches first one will recognize the individual actors, second approach recognize the multiword feature and third one recognize the better names or concept from document. The approach discuss auto slag is to extract Part of speech (POS). Plaintiff algorithm use to extract the synonyms multi words. Similarly in third approach plaintiff will found names on regular expression. Then auto slag will improve the names extraction. This process improves the text representation and concept by giving multiple options.

3.12 A Novel Use of Statistical Parsing to Extract Information from Text [29]

Since 1995, a statistical parsing algorithm has breakthrough in parsing accuracy which was measured against the UPenn TREEBANK. In this paper, Scott Miller, Heidi Fox, Lance Ramshaw, and Ralph Weischedel are adapting a lexicalized, probabilistic context-free parser for information extraction. This technique can evaluate MUC-7 template elements.

This paper highlights statistical approach for parsing information using partial probability model. At each step in the process, a choice is made from a statistical distribution, with the probability of each possible selection dependent on particular features of previously

generated elements. The partial probability over word may be term/component here component means Noun, verb and adjectives. But it is tree based approach not graph based approach. There is need of graph based model because domain is representing association among concepts and the clique. Hence this model is appropriate from graphical prospective.

3.13 Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching [31]

Many important problems involve clustering large datasets. Although naive implementations of clustering are computationally expensive, there are established efficient techniques for clustering when the dataset has either (1) a limited number of clusters, (2) a low feature dimensionality, or (3) a small number of data points. However, there is little work on methods for efficiency of clustering datasets. Andrew McCallum, Kamal Nigamy Lyle H. Ungar present a new technique for clustering these large, high dimensional datasets. The key idea involves using a cheap, approximate distance measure to efficiently divide the data into overlapping subsets called *canopies*. Then clustering will be performed by measuring exact distance between points that found canopy. The canopies in clustering problems were formally impossible. Canopies can be applied to many domains and used with a variety of clustering approaches like Greedy Agglomerative Clustering, K-means and Expectation-Maximization. Andrew McCallum, Kamal Nigamy Lyle H. Ungar present experimental results on grouping bibliographic citations from reference section. Here the canopy approach reduces computation time over a traditional clustering approach by order of magnitude and decreases error in comparison algorithm by 25%.

But disadvantage of this approach is inefficient clustering where data extracted from web because of features about canopies for similar data will result to separate clusters. Feature are not millions for data rather few for special data like protein names. But when approach uses for patient data our new approach of similarity measure will produce better results.

3.14 Learning to Combine Trained Distance Metrics for Duplicate Detection in Databases [26]

Duplicate detection is an important problem in “data cleaning,” and an adaptive approach that learns to identify duplicate records for a specific domain has clear advantages over a static, domain-independent method. Our approach uses learning at two levels. First, similarity metrics are trained to identify duplicate values for each field. Second, multiple similarity metrics for each field are combined to learn a final function for identifying duplicate records. Experimental results demonstrate that this approach detects duplicates more accurately than competing static approaches. In addition, results demonstrate that both levels of adaptation independently contribute to improving the overall accuracy of the system. Real and synthetic datasets show that our method outperforms traditional techniques.

Mikhail Bilenko and Raymond J. Mooney give approach that use two approaches combine to improve results first it will classify / label duplicate / non duplicate record. Then it will identify the duplicate record by similarity metrics and Jaccard approach. But this does not identify the phrase boundaries; pos tags rather identify only duplication. Similarly it is not so efficient algorithm for large dataset. It is only for database domain not for corpus data as it works at record level.

The approach canopies for finding small clusters is work fine for database case but will produce wrong result if it applied to information extraction.

3.15 Automating the Construction of Internet Portals with Machine Learning [25]

The amount of information available on the Internet continues to grow exponentially. As this trend continues, we argue that not only will the public need powerful tools to help them sort through this information, but the creators of these tools will need intelligent techniques to help them build and maintain these services. This paper of Jason Rennie Andrew Kachites McCallum present that machine learning techniques can significantly aid the creation and maintenance of domain-specific portals and search engines. We have presented new research in reinforcement learning, text classification and information extraction towards this end.

In addition to the future work discussed above, we also see many other areas where

machine learning can further automate the construction and maintenance of portals such as ours. For example, text classification can decide which documents on the Web are relevant to the domain. Unsupervised clustering can automatically create a topic hierarchy and generate keywords (Hofmann & Puzicha, 1998; Baker et al., 1999). Citation graph analysis can identify seminal papers (Kleinberg, 1999; Chang et al., 1999). We anticipate developing a suite of many machine learning techniques so that the creation of portals can be accomplished quickly and easily.

This paper classifies the data hierarchically by using labeling technique of Expectation maximization which identify the label and distance label data together. Next it use bayias naïve technique to identify the hierarchies but both approaches have different procedure, input and output. Hence this procedure violates unification rule of thumb. Similarly, the approach is combination of both approaches. This paper also discuss the HMM which is not feed back ward algorithm to identify the maximum likelihood. But this algorithm is not feed forward and feed back ward algorithm.

3.16 Automatic Extraction of Protein Interactions from Scientific Abstracts [17]

Customize the linguistic element, we turned to templates: outline summaries of the information in text. For the information extraction technology to work effectively there must be sufficient detail in the input texts for the contents of the templates to be recognized explicitly. A set of rules (essentially pattern matching rules with a statistical component) for assigning the entities and events recognized by the natural language component to slots in the templates was written and tested.

JamesThomas,David Milward,Christos Ouzunis identify that patterns are identified based on regular expression. Regular expressions are evaluated based on grammar. Grammar is either CFG or regular grammar. Our approach is conventional approach it predict relationships and other patterns but their recall is low. Recall measure on the basis of feature correctness. It is base on classification percentage.

3.17 A Web-resource for Exploring Protein Co-occurrences in MEDLINE Abstracts [18]

It allows identifying the set of Medline abstracts associated with user input, which is considered to be one or more protein names. The input is processed by the ProtScan and proteins identified in the input are used to retrieve the IDs of all MEDLINE abstracts describing the corresponding proteins using protein-to-MEDLINE index. If more than one proteins are identified in the input, the corresponding set of MEDLINE IDs are combined by Boolean 'and' operation. The resulting set of IDs is submitted to PubMed through the published CGI interface to retrieve and display the corresponding abstracts.

Second, the system allows identifying and display a set of proteins associated with user input as well as retrieve MEDLINE abstracts describing each association. This is accomplished in two steps. First, all the MEDLINE abstracts IDs associated with the user input are retrieved as described above. Next, the system retrieves a set of proteins associated with each MEDLINE ID using reverse MEDLINE-to-Protein index and combines them in a single set of unique protein IDs concurrently calculating number of abstracts referencing each protein. Identified proteins are sorted by calculated citation frequency and presented to a user. Clicking on each protein in the found list retrieves a set of abstracts associated with a pair: user input selected protein. This set is calculated on-the-fly as a subset of MEDLINE abstracts simultaneously describing proteins in user input and selected protein.

Above mentioned approach mixed one with dictionary and extract data from text copra. Dictionary based approach is mixed one approach although it produce good result. Our approach is based on single model and clusters. The citation approach measure total results on the basis of cross calculation on MEDLINE. But it does not keep track other features about proteins like synonym, interactions for particular disease type.

Our approach innovate the purpose of unification rather than extraction and citation. This approach also produces different type of protein data from MEDLINE.

3.18 Unsupervised Gene/Protein named Entity Normalization using Automatically Extracted Dictionaries [19]

Aaron M. Cohen developed such algorithm that can handle practically any size input text, in practice the input will usually be individual sentences or abstracts, and this is the input size to which we have tuned our system.

This can be helpful for increasing the sample size for further text mining. For speed and efficiency, first of all this algorithm searches into dictionary. This algorithm does not prohibit sequence of characters bounded by these delimiters. Search is performed based on case sensitivity so it could match case sensitive words.

One of the benefits of the dictionary-based approach is that it is simple and amenable to code optimization. In our case we were able to gain almost a thousand-fold speed improvement over brute force searching against every term in the database. We accomplished this using an approach based on indexing the term prefixes, taking each unique sequence of n initial term characters as the index for all terms with that initial sequence. In our system we chose an n of 6 as a good balance between performance and memory requirements.

Above mentioned approach is a mix one dictionary based. Optimization based approach which uses for better search. Our approach is not dictionary based approach rather than it is unification oriented. This approach is hardly dependent on dictionaries rather than extract data from text corpora. We extract the data from corpus and also increase extraction gradually. The data size and performance concern are not characteristics of this research. We mainly, unify the results of both approach based on different method. It will increase the data mining classification precision.

3.19 A Protein Interaction Extraction System [20]

PIES is protein interaction extraction system developed by Limsoon Wong. PIES obtain all abstracts from Medline those are satisfying the search specification "calyculin". Each abstract is analyzed to identify sentences that mention interaction of proteins, drugs and molecules. Each such sentence is considered an evidence for an interaction. The "actor"

and patient of each interaction are identified. These interaction evidence sentences are then grouped by actor and patient. The result is then displayed in textual form.

PIES put interactions together into a pathway diagram where the nodes are the proteins and the directed edge denote interactions (green arrows for activate and red arrows for inhibit) between the connected protein.

PIES is a system shows the interaction in the graphical form. The resulting graphical representation is clearly shows the output extracted from Medline abstract. The graphical representation extracted from online web site pertaining to Medline abstract.

In contrast to PIES approach we display data into different partition i.e. k partition. The resulting data extracted shows upon query from database into table. Similarly we used matrix for both type of representations.

3.20 Automatic Extraction of Gene and Protein Synonyms from MEDLINE and Journal Articles [21]

Hong Yu, Vasileios Hatzivassiloglou, Carol Friedman, Andrey Rzhetsky, W. John Wilbur developed system in which genes and proteins are often associated with multiple names, and more names are added as new functional or structural information is discovered. Because authors often alternate between these synonyms, information retrieval and extraction benefits from identifying these synonymous names. We have developed a method to extract automatically synonymous gene and protein names from MEDLINE and journal articles. We first identified patterns authors use to list synonymous gene and protein names. We developed SGPE (for synonym extraction of gene and protein names), a software program that recognizes the patterns and extracts from MEDLINE abstracts and full-text journal articles candidate synonymous terms. SGPE then applies a sequence of filters that automatically screen out those terms that are not gene and protein names. We evaluated our method to have an overall precision of 71% on both MEDLINE and journal articles, and 90% precision on the more suitable full-text articles alone.

TH-4320

Automatic extraction of gene and protein synonyms used dictionary based approach and extraction from Medline as well. They use synonym patterns to extract concept from database and from corpus. We inspired from their patterns technique as their patterns are healthier to extract exact proteins. Our approach use patterns for morphological similarity but this approach is not autonomous. This approach is not use for unification to produce useful features and classify the data.

3.21 Comparison Summary

The Support vector machine technique utilize [2] for classification of patterns to recognize either protein. But it does not recognize other features about protein. Similarly, SVM classify data on vector function but it does not completely identify the proteins cross references, abbreviations of protein. SVM does not classify data into different parent categories like cancer, non cancer.

Baysian is base on directed graph approach the primary purpose of baysian model is to predict result rather classify. But their procedure is to identify the partial probability but depend on prior probability. The problem domain does not have prior and posterior probabilities.

K-mean algorithm classifies data into cluster based on numeric features. The model classifies it on distance from center but it does not care some of words which are close to center point. Similarly k-mean is useful if cluster sparse all over the edges but our problem final cluster is dense.

k-mode algorithm is use for dense clustering problem and our approach chooses k-mode for unification. K-mean and k-mode complexity is more than our approach as there is few iteration.

Chapter 4

Methodology

4 Methodology

This chapter includes the methodology which has been used in this research. Methodology has been briefly discussed with respect to changes applied in it. Due to application of these changes, a new methodology is developed and explained as under:

4.1 Prior Approaches

Following are the prior approaches discussed:

4.1.1 Machine Learning technique

These are potentially potent problems in protein name because they affect overall output. Proper nouns missed in many sentences. Complex names "CITR LIGAND,C-Cbl" may confuse the process of extraction during processing of existing techniques (like boosted wrapper induction, support vector machine). Boosted wrapper induction (BWI) concerned with predefined rules and lack feature based usage.

Frame based technique does not properly identified structure into one parse tree. Many parse trees miss many relationships, which could be identified with single parse tree. Protein names in one problem may produces good result but in other case problems of cross-reference, cross resolution impede extraction process.

Support vector machine (SVM) usually deals to extract protein and its interaction. Support vector machine (SVM) not consider those protein names, which are not present at hyper plane. Hyper plane deals data under plane and not formed a clustered behavior. Automatic rule induction deals with automatic rule induction. Protein data not depend on existing semantics. New semantics differ from old semantics therefore new rule emerge. This problem not consider for cross-reference because feature provisioning not present. Feature such as prefix, suffix, marks, plural characters and length of characters not considered.

4.1.2 Probabilistic Models

In probabilistic models soft mining techniques such as working of association mining, clustering unification with information extraction not truly flexible. There is loose integration among techniques. Constrained Random Field (CRF) of McCallum is not truly integrated with data mining technique rather this is motivation.

This technique lacks Uniform compatibility and flexibility hence unable to accomplished task completely with model properties for unification. Model properties are same output, same outcomes, same parameters and same inference not followed properly but partially. Conventional technique of “A Note on the unification of information extraction and data mining using Conditional-Probability, Relational Models, Amherse , MA 01003 USA, 2003” also proposed a framework which deal both type of data but lacks flexibility of different techniques interpretability, problem domain, results and realization of process in the form of algorithm [1]. It is a motivation process infect. Unification succumbs of five milestones as above mention as model inference.

4.1.3 Generative Models

Harlov Markov Model (HMM) Models are generative models because of their generative nature. They generate all paths against all inputs and extract only maximum path from graph. They generate path based on state, transition weight or strength. This technique does not extract complex names prefix, suffix as lack feed forward-backward nature.

All states representing corresponding observed value to measure against hidden value. A state only passes a transition to other state. A transition plus state weight added together to form maximum path. It is interested in finding of maximum path therefore acyclic nature of graph truncate interdependence.

4.1.4 Prior Agent Model

Main purpose of this expert system is to place between the data and models. Its basic function is to integrate two components in intelligent manner (JTC TENG ET AL. 1998). IEEE proposes this system. In Figure 4.1, system is high-level abstraction of process of extraction only deal natural language processing unidirectional.

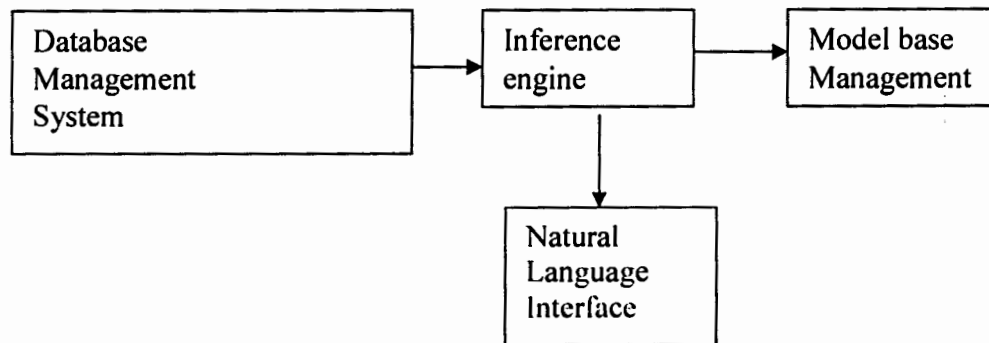


Figure 4.1: Intelligent Model

There are many other agent based architecture deal to extract concept from sentences but not properly unified for both information extraction and data mining. They lack unified admiration of few model properties.

4.1.5 K-Mean Algorithm

Clustering algorithms use for partition, hierarchical, density-based, grid based and model-based methods. Here, we utilize k-mean algorithm for partition. K-mean algorithm further categorizes into details. Partition method of database is base on n objects or data tuples, a partitioning method constructs k partitions of data, where each partition represents a cluster (k less than and equal to n). It classifies the data into k groups, which together satisfy the following requirements:

- Each group must contain at least one object.
- Each object must belong to exactly one group.

Given k , the number of partitions to construct, a partitioning method creates an initial partitioning. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. The general criteria of good partitioning is to relate objects in the same cluster, where some objects are far part in same clusters. There are various kinds of criterion for judging the quality of partitions [25].

To achieve global optimality in the partitioning-based clustering would require the exhaustive enumeration of all the possible partitions. Instead, most applications adopt one of two popular heuristic methods: (1) the k-mean algorithm, where each cluster is represented by the mean value of the objects in the cluster and (2) the medoids algorithm, where each cluster is represented by one of the objects located near centre of cluster. We use mean square algorithm. We use the basic steps of conventional algorithm but clustering based on mean not on minimum distance. We use to calculate the standard deviation [25].

4.2 Our Approaches

4.2.1 Agent Architecture and Messaging

Our architecture communicates using parallel approach. Its design is slightly different because it communicates back from any level to any predecessor or successor. The approach of this model is similar to open architecture [4].

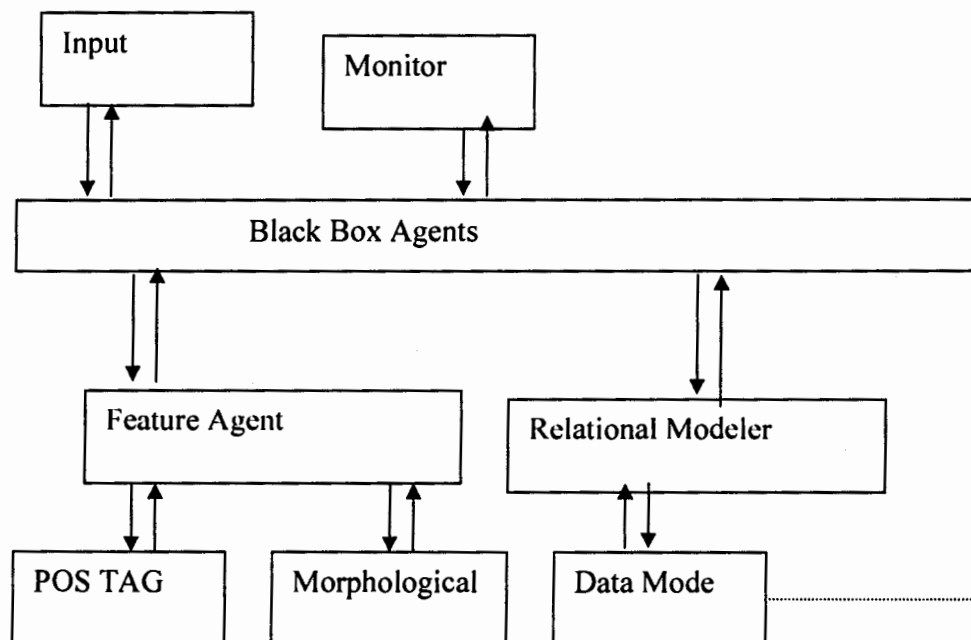


Figure 4.2: Our General Unification architecture using agent for domains Information extraction and data mining

The above model illustrates the parallel and hierarchal communication between different agents. The parallel communication is initiated between children. The hieratical communication will initiate for successor or sub levels. Parallel communicate will depend type of request. Request from remote will treat by monitor agent.

4.2.2 Detail Architecture

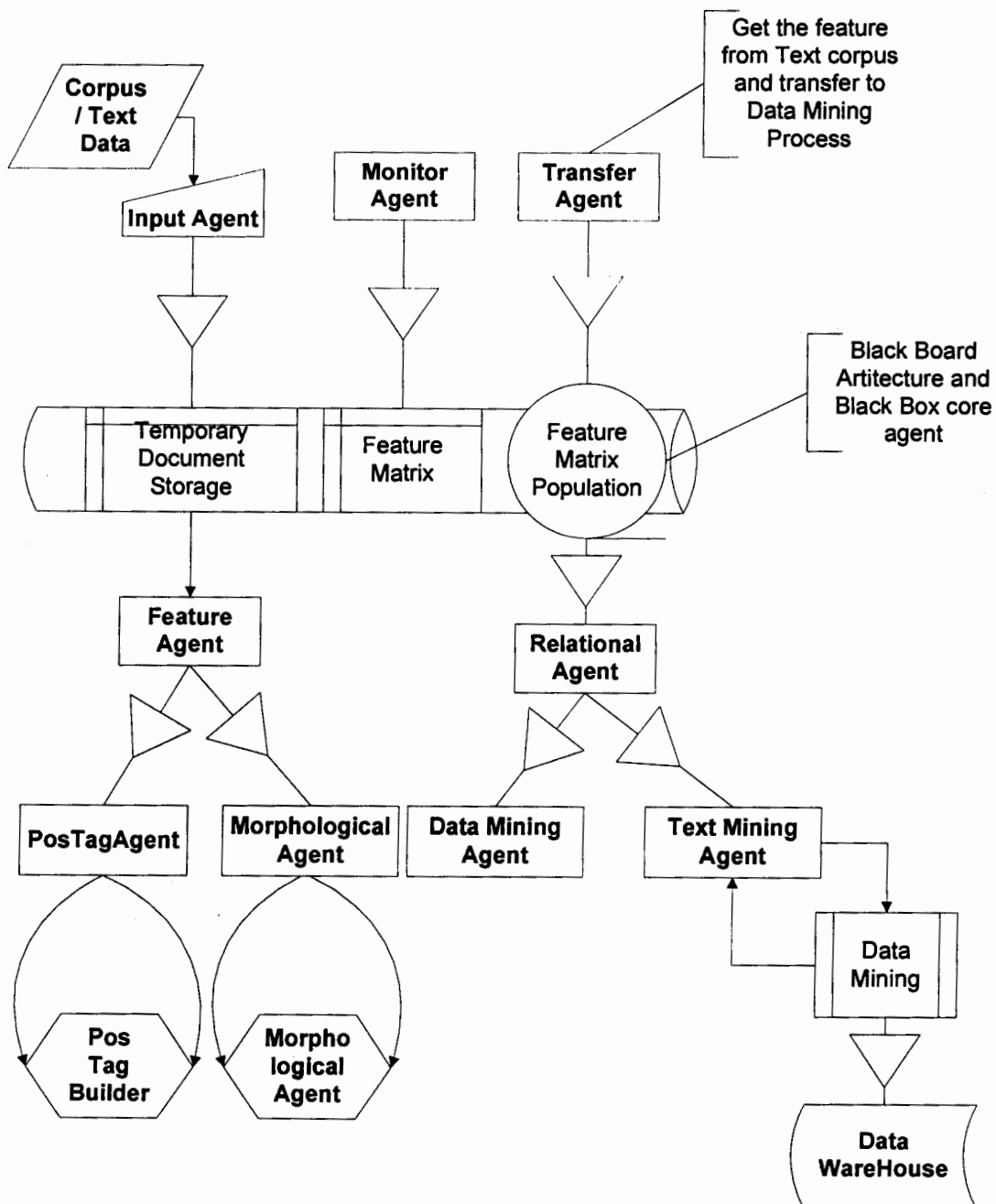


Figure 4.3: Detail unification architecture for information extraction and data mining

An agent could send performative to other agents based on request in Figure 4.3. Any request could of type document, receive features after building and documents send after classification. The feature agent will get request from black box agent which is central service for any new request and black box control the request.

The input will be temporarily store into black box agent. The black box agent will communicate to the respective feature type agent. The feature type agent is ancestor of Pos tag and Morphological tag. The parallel communicate will be initiated from feature agent. Feature agent will send request to morphological agent and pass result back to pos agent. Pos agent will re-initialize feature matrix from feature structure. It will also handle part speech tags for labels noun, verb. It will store the featured structure into the matrix. This matrix stored at black box agent temporarily. The agents like monitoring agent validate the features matrix for validity of features for each entity in Figure 4.3.

Finally, the classification process will start at modeler agent. The request will be handled by modeler agent. The modeler agent will use the feature matrix as similarity matrix. The modeler agent will pass similarity matrix to respective data mining agent or text mining agent. Data mining agent will load the matrix with data feature from database. Next text mining agent will get complete feature list and will pass it to classification algorithm k-mean. The results will be sent back to relational agent, which will pass it to black box agent to display results in Figure 4.3.

4.2.3 Unification Procedure

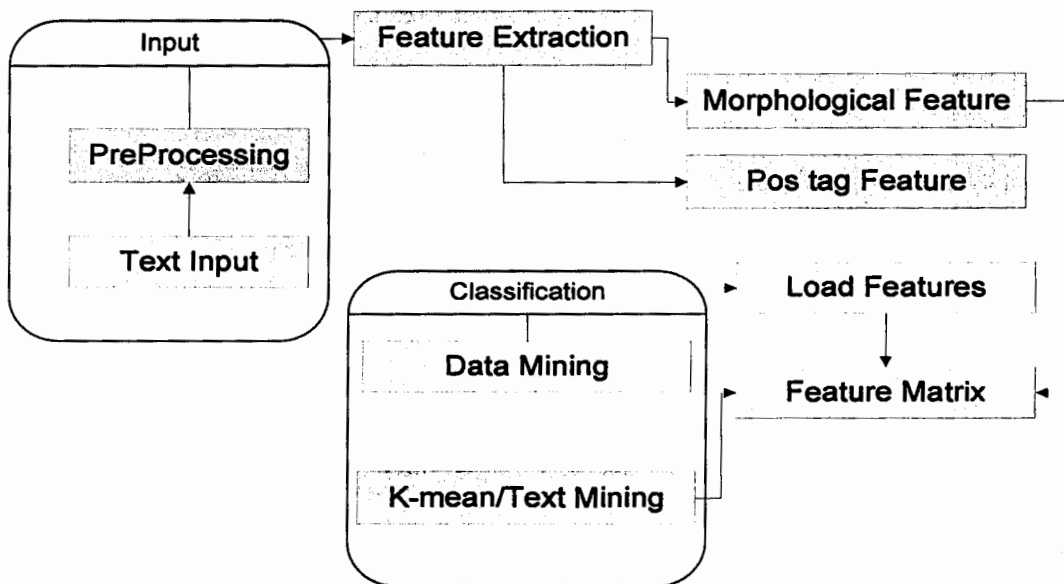


Figure 4.4: Unification Procedure

In preprocessing phase the document will be inputs which are in text format. The document type will also be extracted to predict the whole document domain. Here document means abstract body and its title. During preprocessing title and body will be extracted. The document is about three domains Cancer, HIV and general protein. Next phase is extraction of feature about Cancer, HIV and general protein will be extracted. The feature can be load from data base before text mining phase start for last updated extractions in Figure 4.4. Next the new k-mean algorithm start mining after getting complete sentence and record from corpus and database. All type of feature already measure into matrix like morphological and pos features.

4.2.4 Hybrid Approach of K-mean and Jaccard Approach

Algorithm: k-means. The k-means algorithm for partitioning is base on mean value of objects inside cluster.

Input: The number of clusters k and a database containing n objects.

Output: A set of k clusters that minimizes the squared-error criterion.

Method:

Arbitrarily choose k objects as the initial cluster centers.

Repeat

(re) assign each object to the cluster to which the object is the most similar using min distance, based on the mean value of the objects in the clusters

Update the cluster means i.e. calculate the mean value of the objects for each cluster;

Until no change;

This cluster forms mean value of object in the cluster. The k-mean algorithm takes the input parameter k and partition the n objects into k clusters, hence result of intra-cluster similarity is high but the inter-cluster similarity will be low [25].

K-mean randomly selects k objects, each of which initially represents a cluster mean. For each of the remaining objects, an object is assigned to the cluster most similar basis. The distance between the object and the cluster mean will measure. It then computes the new mean for each cluster. Thus process iterates until the criterion function converges. Converging process is little risky because of cluster boundary compactions [25].

4.2.5 Jaccard Approach

Here we combine two approaches to form the unification for Information extraction and data mining. Jaccard approach measure binary features that were extracted or either load from database. The feature measures by Jaccard approach for similarity of object will be part of k-mean algorithm. Here similarity will be measure by Jaccard approach rather than based on distance because prior probability is unknown in case of corpus raw data.

Comparatively Jaccard approach is non variant similarity based approach. The negative matches are unimportant.

$$D(I, j) = (r + s / q + r + s)$$

This formula shows that r value for asymmetric variables vary as all other values. A binary variable is symmetric if both of its states variable and carry the same weight; There is no preference on which outcomes should marked 0 or 1. Similarity it is base on

symmetric variables, which is called invariant similarity (the result which does not change when some or all of binary variables are coded differently).

Here q is the number of variables that equal 1 for both objects i and j . r is the number of variables that equal 1 for object i but that are 0 for object j . s is the number of variables that equal 0 for object i but equal 1 for object j . Following is measuring dissimilarity from feature matrix which was extracted from text corpus.

Cluster	Interactor	Interactee	Disease	Cancer	Other disease	Synonyms
Object i	1	0	1	1	0	0
Mean 1	1	0	1	1	0	1
Mean 3	0	0	1	0	1	0

Here mean1 means mean of cluster 1 and mean 3 means cluster 3 mean.

$$D(\text{Object } i, \text{Mean } 1) = 0 + 1/3 + 0 + 1 = 0.25$$

$$D(\text{Object } i, \text{Mean } 3) = 1 + 1/1 + 1 + 1 = 0.67$$

Clearly seems that object i is similar with cluster 1 mean so it is will be assigned to cluster 1 and mean will be recalculate. Both approaches have clear distinction in usage. One approach is used for minimizing error square mean but other used to measure similarity value.

4.2.6 New K-Mean Algorithm

We start with three clusters k is equal to 3. These clusters represent different categories of interest. First category is about cancer proteins, second category represents disease specific proteins and third category represents general proteins. First cluster will show result of all interactors, interactee, and synonyms together. In second cluster will show HIV proteins of all. Third cluster will show all other proteins.

New algorithm represent these concept on the based of deviation value. Actually, we analyze some of interactors, interactee. For clustering interactors and interactee, there is need of proper requirement of partitions to classify data. We choose common place to start. Its purpose is to unify, which is core point for different approaches of information extraction and data mining. We define one procedure for all type of data mining techniques and information extraction.

The algorithm will unify two domains on common feature matrix. The data feature will be store into matrix from source database as well. The feature builder agents will populate the matrix before classification initiated for both sources.

Conventional k-mean algorithm measures similarity on minimum distance. Minimum distance calculates from arbitrary point treated as center of cluster. This algorithm calculates distance based on similarity and likelihood of Jaccard approach.

Conventional algorithm is base on similarity matrix but matrix structure is unique. The similarity will be measured on the basis of most similar interactors, interactee and synonym. The similarity is measure on Jaccard approach. Most similar interactors, interactee and synonym accumulate into cancer, HIV and general proteins.

New Algorithm: k-means. The k-means algorithm for partitioning is base on mean value of objects inside cluster.

Input: The number of clusters k and a database containing n objects and n object features.

Output: A set of k clusters that minimizes the squared-error criterion.

Method:

Arbitrarily choose k objects as the initial cluster centers.

Choose arbitrarily after feature extraction at feature agent

repeat

(re) assign each object to the cluster to which the object is the

most similar using jaccard approach, based on the mean value of the objects in the clusters

Update the cluster means i.e. calculate the mean value of the objects for each cluster;

Until no change;

4.3 Cyclicity of Data

Cycle of clusters formed after step by process.

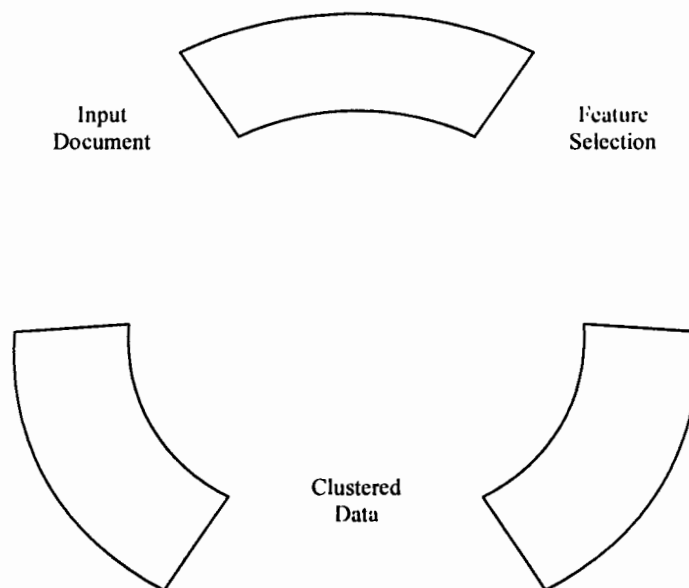


Figure 4.5: Our cycle for formulation of clusters

In Figure 4.5 process of cycle formulation repeated for every next input. First iteration in system is of input document, second iteration is of feature type and last iteration is of clustered data.

At second iteration data's feature will be gathered base on input. Features are in the form of matrix and store temporally into agents. Next last iteration starts for cluster formation and unification point. Last iteration is also serving as unification point for extraction

process and data mining. For data mining K-mean algorithm serve at here based on common outcome and procedure.

Common outcome and procedure starts from first iteration to last iteration based on feature matrix. Classification process will be unifying on feature matrix at third iteration. The classification could be done and new rule can be build for extraction of new features. But it is beyond our scope. Here we build feature matrix mechanism and new classification procedure.

4.4 Infrastructure of Application

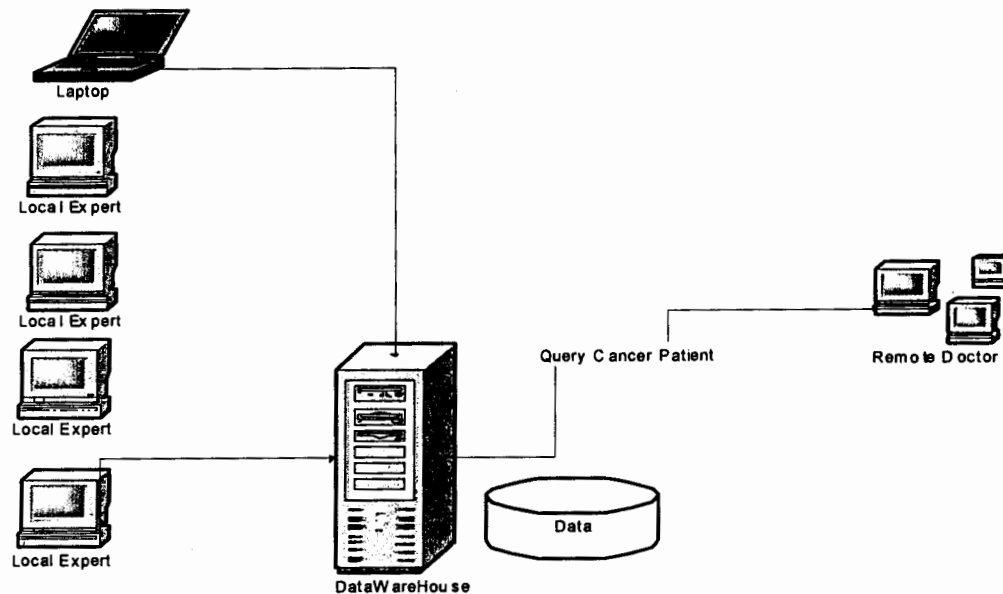


Figure 4.6: Our System Architecture for Information Extraction and Data Mining

In Figure 4.6 architecture illustrates local expert and remote expert doctors. The system agent communicates with database for query and classification of results. If doctor wants information about cancer related patient then doctor will pass query from remote location. Remote query pass to invoke classification process at server site. In detail data modeler will pass request by monitor agent. The monitor agent will communicate to server to monitor agent. The feature extraction and classification process will be initiated at server site. Then all result will be submitted to remote doctor by agent communication.

4.5 Algorithm/ Pseudo Code

Procedure classificationBuilder

Input: Document dn, Sentence Sn

Output: cluster 1, cluster 2, cluster 3

Select from Sn, initial cluster c1, cluster c2, cluster c3 center

1.1 Loop for 1..... Sn / clustercounter

dissimilarity1=Jaccard (c1 center, Si) dissimilarity

dissimilarity2=Jaccard (c2 center, Si) dissimilarity

dissimilarity3=Jaccard (c3 center, Si) dissimilarity

found =: false

If dissimilarity <:= dissimilarity2 and dissimilarity1 <:= dissimilarity3 and clustercounter
=: 1 then

Position 1: = Find in c1 (Si)

Position 2: = Find in c2 (Si)

Position 3: = Find in c3 (Si)

If (position1 > 0) i.e. it found

c1.item(Si) already assigned

End

Else Skip

Else If (position2 > 0) i.e. it found

c2.removeitem (Si)

c1.item (Si) (re) assigned

found =: true

End

Else If (position3 > 0) i.e. it found

c3.removeitem (Si)

c1.item (Si) (re) assigned

found =: true

End

c1 center =: CalculateMean (c1)

```

        If (S:= n and clustercountter1) then
            S i =0 reinitialize loop do again 1.1
        End
    End
End
If ( found !=: true and clustercount==1 )
    Position 1: = Find in c1 (Si)
    If (Position1 >= 0)
        c1.removeitem ( Si )
    End
    Else
    Position 2: = Find in c2 (Si)
    Position 3: = Find in c3 (Si)
    If (Position2>= 0)
        C2.moveitem ( Si )
    End
    If (Position3>= 0)
        C3.moveitem ( Si )
    End
        c1 center  = : CalculateMean (c1)
        If (S:= n and clustercountter1) then
            S i =0 reinitialize loop do again 1.1
        End
    End
End
If dissimilarity2 <:= dissimilarity1 and dissimilarity2 <:= dissimilarity3 and
clustercounter =: 2 then
    Position 1: = Find in c1 (Si)
    Position 2: = Find in c2 (Si)
    Position 3: = Find in c3 (Si)
        If (position2 > 0) i.e. it found

```

```

        C2.item( Si ) already assigned
    End
    Else Skip
        Else If (position1 > 0) i.e. it found
            C1.removeitem ( Si )
            C2.item ( Si ) (re) assigned
            found =: true
        End
        Else If (position3 > 0) i.e. it found
            c3.removeitem ( Si )
            c2.item ( Si ) (re) assigned
            found =: true
        End
        C2 center =: CalculateMean (c2)
        If (S:= n and clustercountter1) then
            S i =0 reinitialize loop do again 1.1
        End
    End
End
If ( found !=: true and clustercount==2 )
    Position 2: = Find in c2 (Si)
    If (Position2 >= 0)
        C2.removeitem ( Si )
    End
    Else
        Position 1: = Find in c1 (Si)
        Position 3: = Find in c3 (Si)
        If (Position1 >= 0)
            C1.moveitem ( Si )
        End
        If (Position3 >= 0)
            C3.moveitem ( Si )

```

```

End
    C2 center  = : CalculateMean (c2)
    If (S:= n and clustercountter2) than
        S i =0 reinitialize loop do again 1.1
    End
End
If dissimilarity3 <:= dissimilarity1 and dissimilarity3 <:= dissimilarity2 and
clustercounter =: 3 then
    Position 1: = Find in c1 (Si)
    Position 2: = Find in c2 (Si)
    Position 3: = Find in c3 (Si)
    If (position3 > 0) i.e. it found
        C3.item( Si ) already assigned
    End
    Else Skip
        Else If (position1 > 0) i.e. it found
            C1.removeitem ( Si )
            C3.item ( Si ) (re) assigned
            found =: true
        End
        Else If (position2 > 0) i.e. it found
            C2.removeitem ( Si )
            c3.item ( Si ) (re) assigned
            found =: true
        End
        C3 center  = : CalculateMean (c3)
        If (S:= n and clustercountter3) than
            S i =0 reinitialize loop do again 1.1
        End
    End
End
If ( found !=: true and clustercount==3 )

```



```

    Position 3: = Find in c3 (Si)
    If (Position3 >= 0)
        C3.removeitem ( Si )
    End
    Else
    Position 2: = Find in c1 (Si)
    Position 3: = Find in c3 (Si)
    If (Position2>= 0)
        C2.moveitem ( Si )
    End
    If (Position1>= 0)
        C1.moveitem ( Si )
    End
        C3 center  = : CalculateMean (c3)
        If (S:= n and clustercountter3) than
            S i =0 reinitialize loop do again 1.1
        End
    End
    End
    If (clustercount = 4)
        Iteration++
        If (iteration=4)
            Clustercountcount=4
        End
        Clustercountcount=1
    End
    End
    Loop again at 1.1
    Procedure classificationBuilder end
    Procedure Jaccard (c1 center/mean,S1)
    Q := 0, R:= 0, S:=0
    Loop
        If (c1.var == 1 and S1.var ==1)

```

```
Q++
Else If (c1.var == 1 and S1.var == 0)
R++
Else If (c1.var == 0 and S1.var == 1)
S++
End loop
Return r+s / q + r + s
Procedure Jaccard end
```

Chapter 5

Implementation / Results

5 Implementation / Results

This chapter covers all the implementation aspects and their results captured thereafter.

5.1 Application

Database design and physical issues are discussed. Similarly class implementation and design elaborated in detail.

5.1.1 Star Schema

The star schema is composed of fact table and dimensions which are cancer Type, disease, interactee, interactors, disease and general type. The fact table is subject about which all interactions will be recorded. The user wants to calculate total interactors, total abbreviation, total synonyms and total cancer summaries over particular type of disease that is cancer/HIV.

The interactors' compile summary about total abbreviations, total synonyms and total disease from Medline abstracts. The fact table records all findings which were found with interactor in Medline abstract. The abbreviation occurs for any protein name Interactor) i.e. any short name found with period of time.

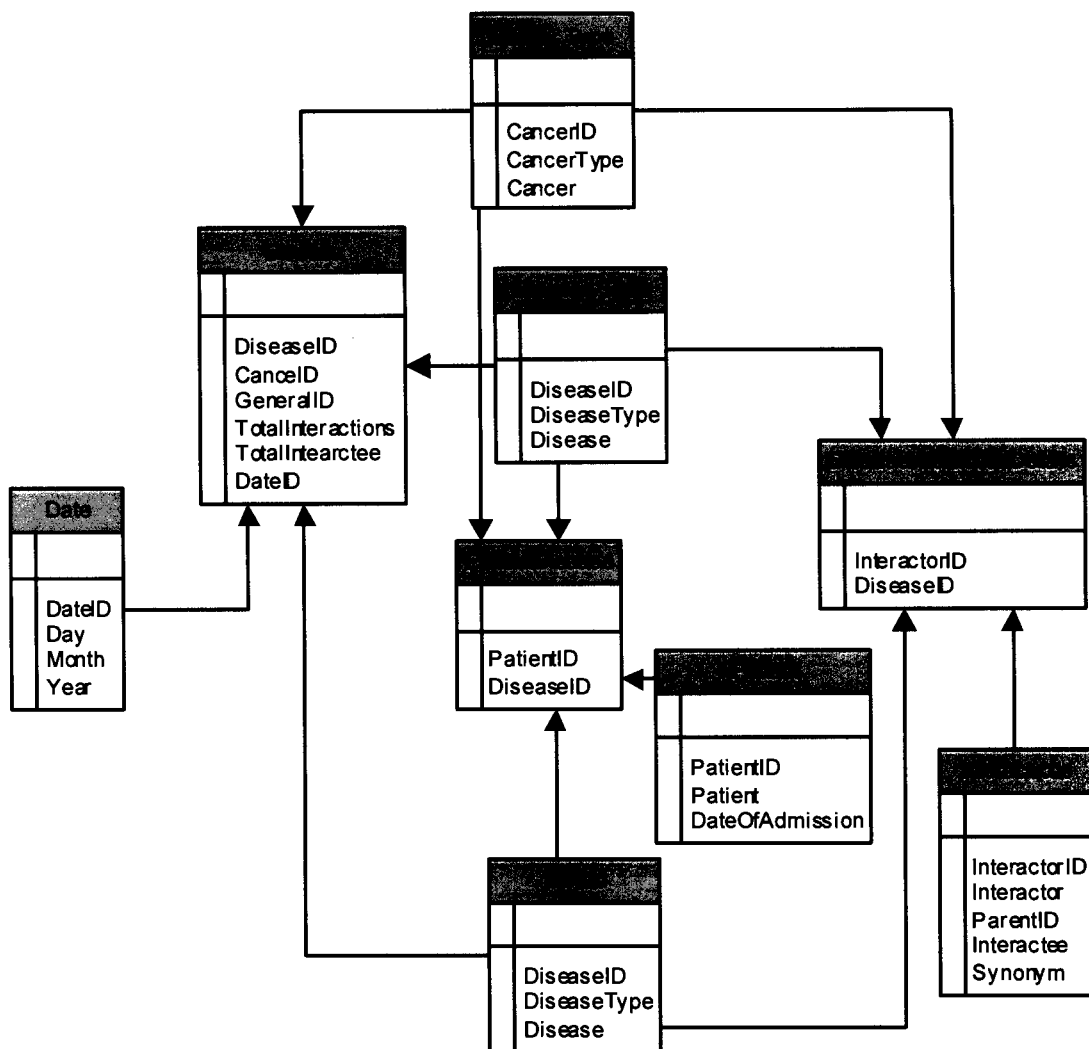


Figure 5.1: Logical Diagram of Protein and Its Interaction

The fact table is filled with numeric entries from Medline copra. The numeric entries are total diseases, total general interactions, total synonyms and total abbreviations. The total diseases are those which calculate on disease occurred in the Medline abstract.

The total abbreviations' are those which calculate for short named proteins. The general interactions are general one. The interactions that are general one which are specific to species. The synonyms are findings which were proteins infect but referring to protein.

The activities are disease type or general types and cancer type recorded against fact table. We can find particular disease and its involving protein and its interactions. The disease type activity occurs due to new disease finding and new interactions regarding disease.

The dimensions are interactions and disease and documents. The interaction and disease records are in dimensions at source site and updated periodically. The disease type predict the overall disease if occurs in interactions. The date and time dimension information is used to calculate periodic information.

The query of doctor against any patient disease will be initiated i.e. doctor wants smith blood cancer related latest protein details. Our warehouse kept about 1999, 2000, and 2001 updated dimension about disease. The response of query will be in the form of interactee and interactor which are protein infect associate to cancer disease.

5.1.2 Transformation

The layers are divided into three parts. First layer is about source like Medline abstract, database and raw dictionaries. The second layer communicate to first layer while transformation occur with batch Loads. The loading of data from sources occurs in batches. The interactions are loaded with all other dimensions and fact table. The major factor of loading of document is base on at least ten documents.

The communication from source to warehouse is base on direct warehouse. The last layer is used for application or data mining purpose. Here, we unify information extraction with data mining after getting data from data warehouse. The transformation occurs on batch documents and these batch documents transforms after every ten continuous updates at client.

5.1.3 Volume of Data

The volume of data continues to increase as raw data in the form of medline abstract grow. The volume of data first brings in the form of documents then transform into transaction. The important feature will be founded by rules building. The data in warehouse record fact table which also indicates number of year. The volume increase Medline abstract increase on every year.

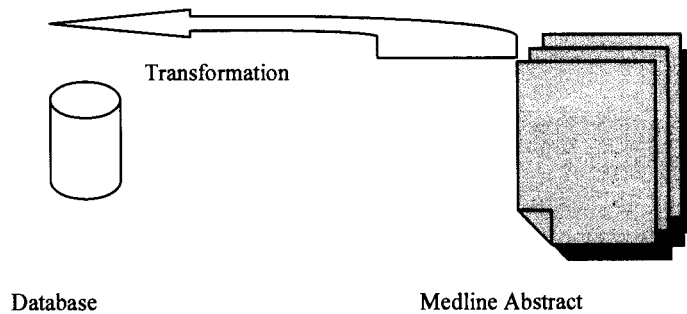


Figure 5.2: Volume of data

5.1.4 Physical Issues

There are following physical issues.

5.1.4.1 Rollover Capability

Analyze interactions i.e. proteins and their reference records for fact Disease. This design has summaries and rollover capability for specific fact data. Roll over summaries will be calculated on days and cumulative on month and year basis.

For example, From May 1 1998 to May 1 1999 will be roll over for particular disease.

5.1.4.2 Frequency of Access

Columns are grouped on frequency of access. Where level of frequency is low that is not concern so they are grouped together into disease type, cancer type and other table. The indexes are use on high frequency table such as Disease fact table.

5.1.4.3 Merging of Tables

This design has capability to merge together with the periodic data. Due to merging of tables there requires less I/O. Merging of dimension table could be possible where loading will be done.

5.1.4.4 Redundancy of Attributes

In design there is requirement of minimum redundancy of attributes so they seldom to use. Our design group repeated attributes in case of date.

5.1.5 Profiling

Separate profiles are created for the purpose of aggregation of data. These profiles summarized data for data mart purpose. Here in this design the data mart are hospital sites.

5.1.6 Operating System Issues

The operating systems LINUX have SQL query naming issues and other development issues in case of querying from application.

5.1.7 Nature of Ware House

The warehouse is of direct nature where an expert can query from application. Query request will be handed by web after wrinkle of time. The indirect nature of warehouse is not feasible because there is intermediate layer who will answer the queries behalf of doctor.

5.1.8 Class Diagram

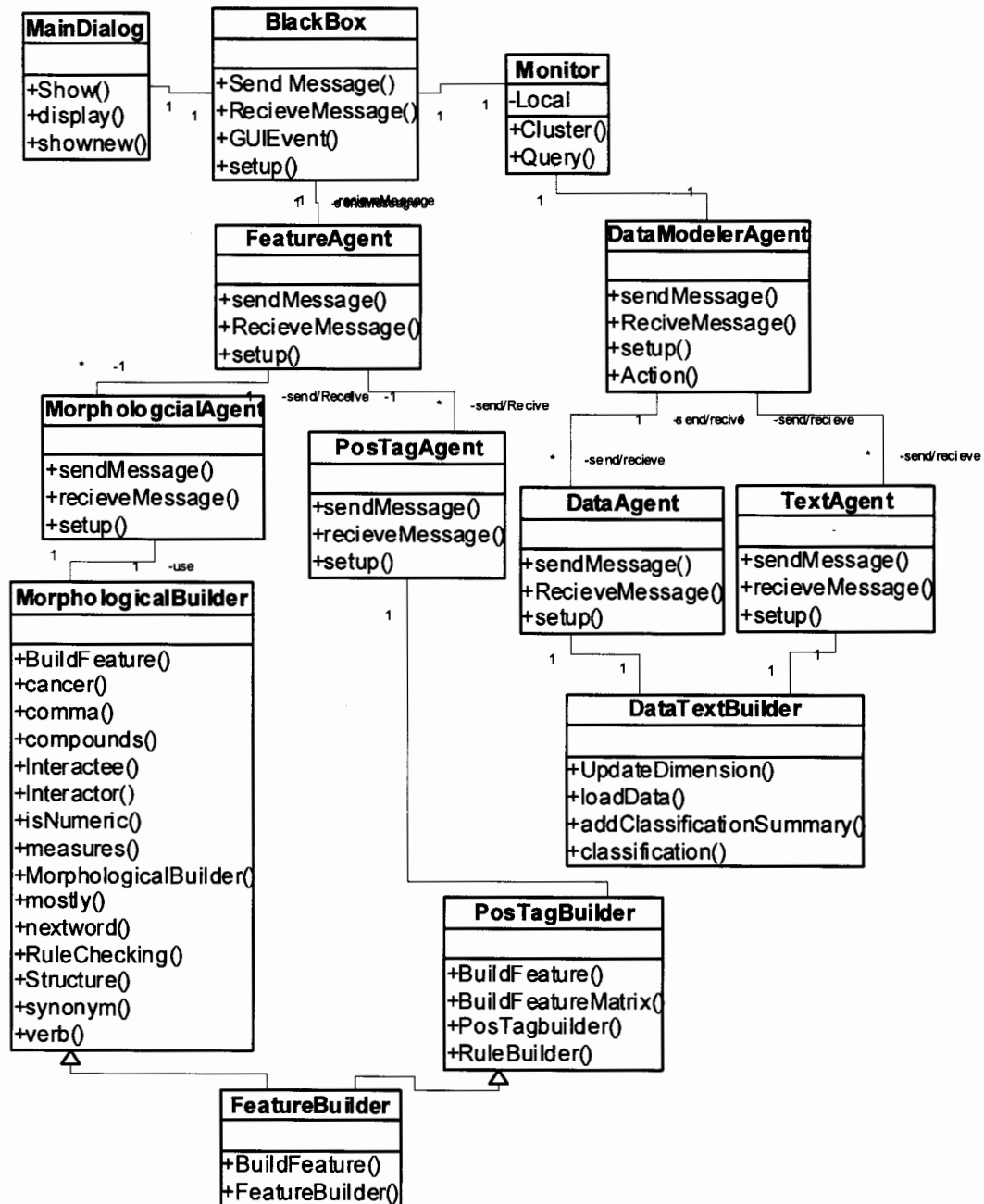


Figure 5.3 Main Class Diagram

Above main diagram shows the main architecture classes. The explanation of each class and method is as follows.

5.1.8.1 Black Box

This class is main and core class of architecture send message by send message () function. Initially black box agent will send message to feature agent the send message will take all documents and its sentence in the form of object.

Receive message () will receive message from feature agent and here the features will build and text will be annotated.

GUI event method is use for GUI relate event handling. Here parameters will be received from dialog. The main dialog will send parameter to Black box GUI method to handle it. It is Model View based architecture between Black Box and Main dialog.

```
List first = ((MainDocument) buildFeature).firstCluster();  
List second = ((MainDocument) buildFeature).SecondCluster();  
List third = ((MainDocument) buildFeature).ThridCluster();  
  
Vector col=new Vector();  
List total = ((MainDocument) buildFeature).getAlltext();  
  
Vector row=dialog1.createTableHeaders(second);  
dialog1.setColHeaders (row , col);
```

Here you can see black box will display clusters after classification into data grid.

5.1.8.2 Feature Agent

This class is responsible for sending and receiving message from morphological and pos agent. The morphological agent will send feature data to feature agent after extracting feature about data. Feature agent will initiate parallel communication and feature matrix will be build from pos agent.

The communicate ends on feature extraction and matrix building. First message will be send to morphological agent for feature building. Second message will send to pos agent for matrix filling. Feature agent will receive two message form pos agent and from morphological agent.

```

if (msg1.INFORM==msg1.getPerformative())
{
    // SEND LIST FOR FEATURE EXTRACTION
    sendMessage("morp",list);
}
else if ( msg1.AGREE==getPerformative())
{
    // SEND POS TAGS for FEATURE MATRIX
    sendMessage("pos",list);
}
else if ( msg1.CONFIRM==getPerformative())
{
    // SEND IT BACK TO BLACK BOX
    setPerformative(msg.QUERY_REF);
    sendMessage("black",list);
}

```

Feature agent will handle parallel message for both of its predecessor's black box and pos agent. Finally, send results back to black box agent.

5.1.8.3 Morphological Agent

This agent will get build feature and call buildfeature () method of morphological agent. All methods cancer, comma, interactee, interactor, synonyms will be called in rule checking function. The function will be called on every word of senetence in buildfeature() method. The final feature extracted will be sending to feature agent.

```

if (msg1.REQUEST==msg1.getPerformative())
{
    build = new MorphologicalBuilder();
    map = build.BuildFeature(list);
}

```

The morphological features will be building at here in morphological builder.

5.1.8.4 POS Tag Agent

This class will fill the feature matrix and annotation feature will be build from morphological features. It will call Buildfeature () function for annotation feature and matrix filling. On every sentence the matrix will be filled with extracted feature and call buildfeaturematrix () function.

```

if (msg1.REQUEST==msg1.getPerformative())
{
    MainDocument mainDoc=posbuilder.BuildFeature((HashMap)list);
    setPerformative (msg.CONFIRM);
    sendMessage ("feature",(MainDocument)mainDoc);
}

```

Part of speech (POS) feature will be build in parallel with morphological features.

5.1.8.5 Monitor Agent

This agent class will initiate communication by checking the local and remote request from attribute local. If request is coming from remote location then it will check the request type either cluster request or query request. Finally, it will send message and request parameter to DataModeler agent. In case of classification request the flag will be false and parameter like feature matrix and structure of document will be sent to datamodeler agent. If request is about query then query parameters like cancer type or other disease pass to datamodeler agent.

5.1.8.6 Data Modeler Agent

This agent will receive feature matrix and request about query or clustering from monitor agent. If request is coming from local then classification request will be initiated. First it will send message on data agent. The data agent will call load data function to load data from database. The structure will be filled back by loaddata () function. The message reply ok will be sending back to message initiator. Data modeler agent will send message back to text agent for clustering. The text agent will send back data after classification.

5.1.8.7 Data Agent

Data agent will receive request from DataModeler agent to update dimension like disease and their interactions data. Data agent will call update dimension method to record information about new Medline abstract for all diseases. Then it load update data on loaddata () function the feature matrix will be rebuild over their and send back to DataModeler agent.

5.1.8.8 Text Agent

This agent will handle new clustering process and will receive message request of clustering after re-building of matrix from data agent. The message of feature matrix will be send to this agent and it will call addclassificationsummary() function. This function will classify the final out put and record fact information in database. It will send message and three classified clustered back to data modeler agent. DataModeler agent will send data back to black box agent to display final output.

```
if (msg!=null&&msg.getPerformative()==msg.DISCONFIRM)
{
    MainDocument doc1=null;
    ClassificationBuilder build=new ClassificationBuilder ();
    doc1=build.Classify (doc);
}
```

This agent will classify the results into database and send back modeler agent to display results at black box.

5.1.9 Jaccard Approach Implementation

Typical Jaccard approach is implementation has shown in the Appendix A.

Where similarity will be measured between first random cluster point which is meantxt1 and object i which is pointText. Then similarly random second cluster and third cluster values will be measured. Finally, Jaccard values will be used as similarity value for cluster rather than distance.

5.1.10 Features Agent Implementation

The feature agent is responsible for the sending the data set for feature measurement to two its two children. One is responsible for the morphological features and other is responsible for Part of speech features. For the code that handle feature measurement in parallel sees Appendix A.





5.1.11 Rule Checking by Feature Agent Ancestor

This method indicates the verb features use for association between protein and its interaction. Rule checking performed on verb and synonym. The rule performed for verb and synonym at separate methods see Appendix A.

5.1.12 Special Method and Feature Building

Numeric checking which is used for elimination of the numeric data and numeric data found at position of noun position. See Appendix A for special feature measurement.

5.2 Collaboration Diagrams

			
Message	Document structure	Cluster output	Feature Matrix

5.2.1 Black Box to Feature Agent

First message will send after loading data from directory and filling structure to feature agent.

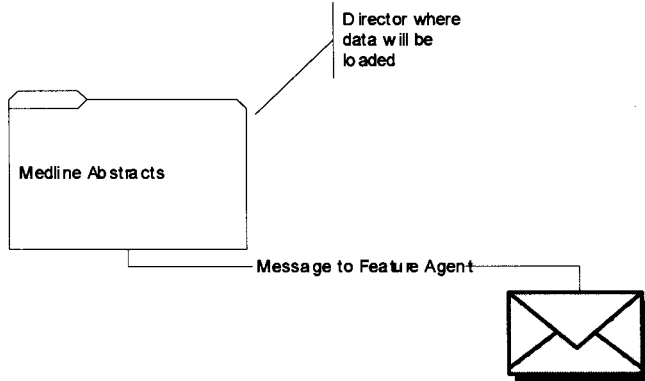


Figure 5.4: Black Box to Feature Agent

5.2.2 Feature Agent to Morphological Agent

Second message will send to morphological agent to extract feature from Medline abstract.

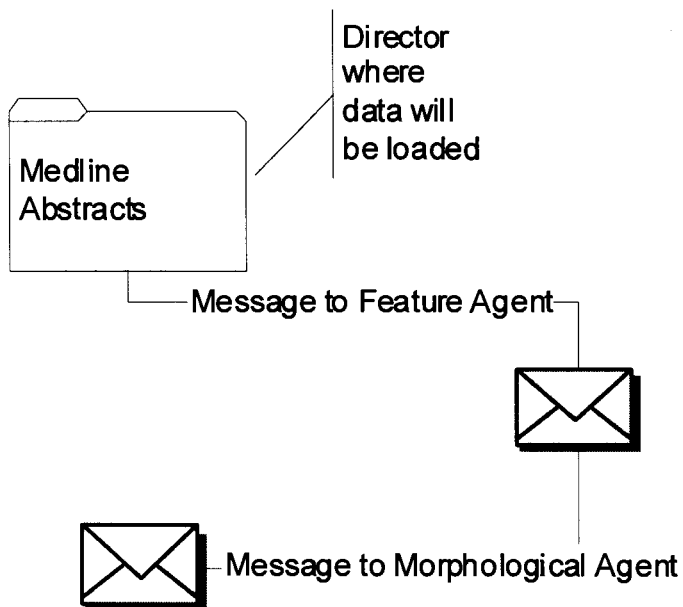


Figure 5.4: Feature Agent to Morphological Agent

5.2.3 Morphological Agent to Feature Agent

Third message will send to back to feature after completion of feature extraction from Medline abstracts.

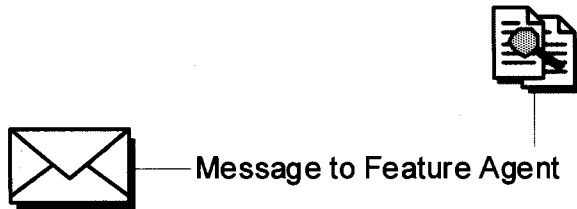


Figure 5.5: Morphological Agent to Feature Agent

5.2.4 Feature Agent to Pos tag Agent

Fourth message will send to pos tag agent for feature matrix building which will use for unification.



Figure 5.6: Feature Agent to Pos tag Agent

5.2.5 Pos tag Agent to Feature Agent

Fifth message will send back to feature agent it but feature and feature matrix will send as well.

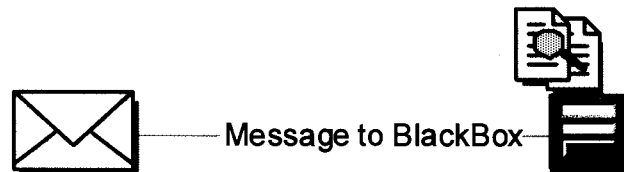


Figure 5.7: Pos tag Agent to Feature agent

5.2.6 Feature Agent to black box

Message will sent back to black box agent.



Figure 5.8: Pos tag Agent to black agent

5.2.7 Monitor Agent to Data Modeler Agent

Sixth message will send to data modeler agent who will handle for remote query request.



Figure 5.9: Monitor agent to data modeler agent

5.2.8 Data Agent to Data Modeler Agent

Seventh message will sent back about query request and data modeler will get it.



Figure 5.10: Data agent to Data Modeler agent

5.2.9 Data Modeler Agent to Black Box Agent

Eighth message will send back results to display on black box.



Figure 5.11: Data Modeler agent to black box

5.2.10 Monitor Agent to Data Modeler Agent

Here fifth message will be considered at 4.7.6. Sixth message will send to data modeler agent, which will be for classification request.



Figure 5.12: Monitor agent to Data Modeler Agent

5.2.11 Data Modeler Agent to Data Agent

Seventh message will send to data agent, which will update the feature extracted and fill feature matrix.



Figure 5.13: Data Modeler Agent to Data Agent

5.2.12 Data Modeler Agent to Text Agent

Eight message will send back to data modeler agent about classification from rebuild feature matrix.



Figure 5.14: Data Modeler Agent to Text Agent

5.3 OS (Operating System) Settings

Specifications for operating system are Linux 4.5 and Oracle 10 g supported at that platform. There need of samba connectivity from Windows NT system to Linux system. Samba configurations are necessary for data loading from sources or data marts to warehouses. Because there are multiplatform so we need samba configuration.

- 1 Multi-platform support at database
- 2 Need of case sensitive SQL and PLSQL

5.4 Batch Processes

The ten documents will be passed from input agent to feature agent for feature extraction.

Repeat until all document read

*The outer loop will continue to used until entire process complete

Repeat until n document read

*The inner loop will continue to used until all documents read

```
int i=1;
```

```
File f=new File ("c:\text"+i);  
i++;  
document. Text=f.read ();
```

Next

*pass documents collection to black box

```
Blackbox black=new blackbox ();  
Black.reposirtory (documents);  
Serialize (Black);
```

*pass documents collection to Feature agent

```
Featureagent agent=new Featureagent ();  
agent.ontolligies (documents);  
agent.send(pos agent);  
agent.send(morphological agent);
```

*Parallel Behavior

```
Posagent posagent=new Posagent ();  
featurematrix matrix=posagent.features(documents);
```

Loop until all document read

This process will continue for ten documents and features extracted for every ten document. The process will complete until all of data will be saved into matrix form.

5.5 Error Handling

The noisy data will be caught in case of missing and noisy findings. The data will log into database by monitoring agent.

Class monitoring agent

```

{
Public void Behaviour()
{
Try{
// traverse the matrix and missing feature
//pass data to modeler agent behaviors
}
Catch
{
// missing feature will be saved as new feature or noisy data.
}
}
State the database changes required for this feature.

```

5.6 Agent Communication

Below are the agents and their action which shows the agent communication.

Agent	Description	Action
Input Agent	The agent is responsible for reading documents from disk.	i) Read 10 documents. ii) Display in Textbox iii) Store at black box agent
Black Box	That will store input and output matrix.	i) store input documents
Feature agent	Pass all data with parallel behavior to all Childs.	i) input text pass to all children
Pos Tag	Apply features to data	ii) apply features to data and annotate the text and made matrix

Morphological Tag	Apply features to data	iii) apply features to data and annotate the text and entry into feature matrix
Monitoring agent	Check the missing data and pass further	i) the data will be passed to data modeler agent
Data Modeler Agent	Pass data to respective agent	i) pass data to data miner agent
Data miner agent	Mining of data into clusters	ii) get data from data modeler iii) mine the data with feature matrix iv) save data into data ware house v) display cluster into grid

Table 5.1: Agent Communication

5.7 Results

We use the Medline abstract of year 2000 from the Medical sites. The medical abstract are about proteins and its interaction. The preprocessing of data occurs but before preprocessing of abstract the result will shown in plain text format.

The Figure 5.16 explain the Medline Abstract in the raw format, the publication was related to disease cancer proteins or its association. As clearly seen in first line that is colorectal cancer disease. The Next lines are about the details of research but surely have protein. The body has different sentences and has protein and its interaction. The sentence may have protein at start or at end of sentence. The sentence could have protein as synonym of protein. The synonyms will be treated by analyzing the special words as pattern to predict the synonym existence. These patterns are special words which is also part of particular sentence. The features of synonym are measured with respect to these special words. The analysis that a synonym could occur in the sentence is as follows. Protein could be synonym in sentence like in the form of word 'its'. The word 'its' is referring to the protein in first sentence. It is two sentences analysis problem because if

first sentence has protein then second defiantly has cross referenced protein. The word cross referenced is 'its'.

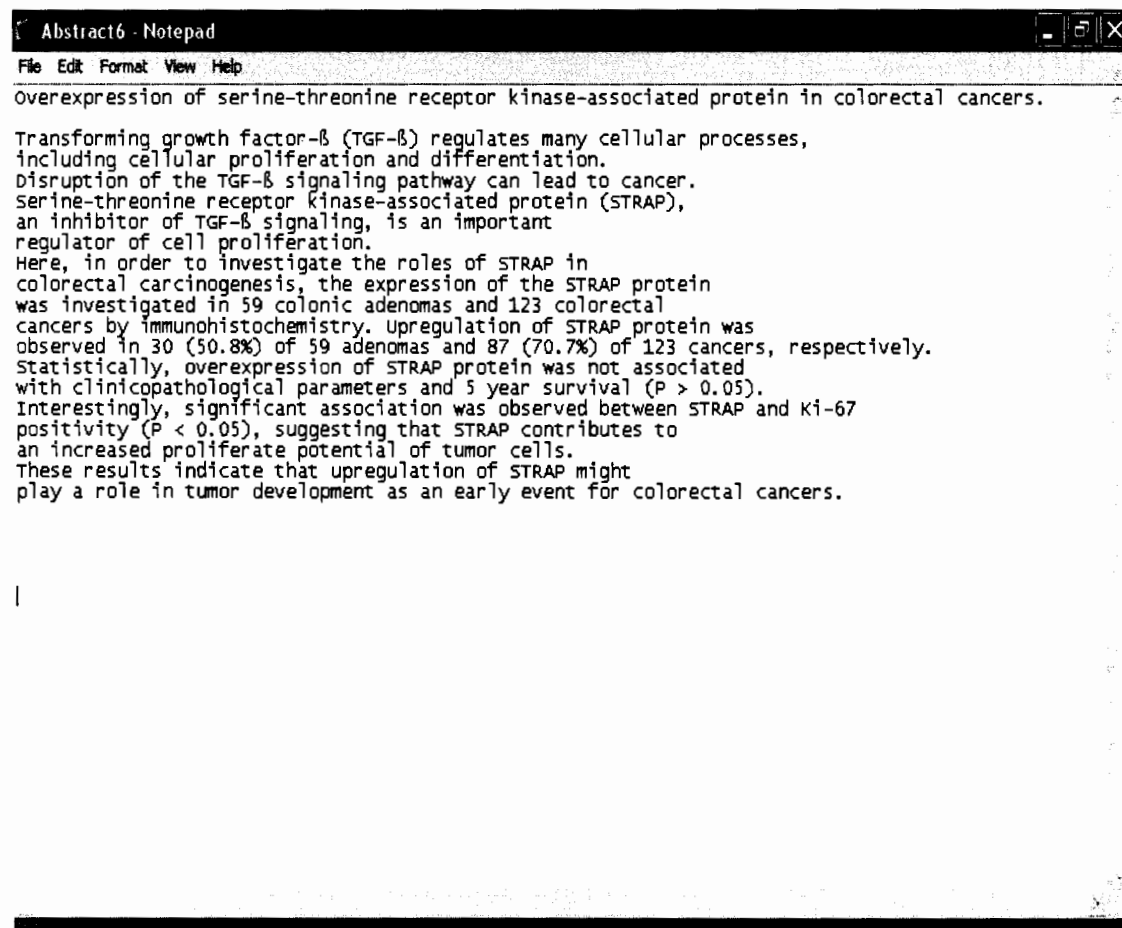


Figure 5.16: Initial format of Medline Abstract

The other type of data is interaction and its interactor problem in sentence. The interactor is protein in the sentence. If the sentence have verb like “compose of , contains” or other type of conventional verbs that occur between one protein (Interactor) and other protein or proteins (Interactee). The verb indicates the interaction type between interactor and interactee.

The sentence could have stand alone protein in its sentence. The sentence could have protein as Noun in the sentence as protein to extract.

5.7.1 Medline Abstracts Main Screen

The sentences will be extracted from the format raw text. The file will be extracted into system. System will extract title and body separately.

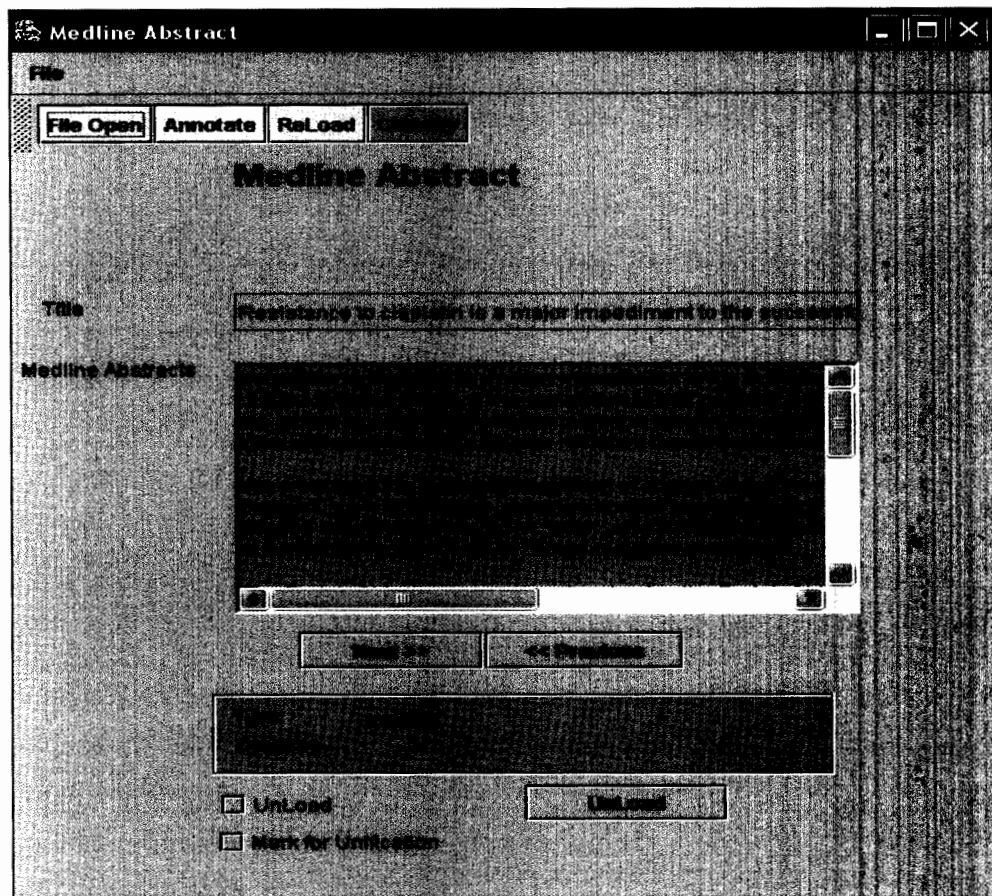


Figure 5.17: Main Screen

The main screen illustrate Medline abstract extracted from corpus database as shown in Figure 5.17. This screen is showing the abstract after applying pre-filtering. The pre-filter will parse the data separately into title and body. Next the basic label will be analyzed that is disease and disease type. The rules for getting disease and its type will be based on

reflection pattern. User need to enter only disease type at one place it will extract automatically. There is flat database of disease and its type is maintained. The word will be match at the time of input from file and match with flat disease database.

Input: D1 (document) to Dn of S1 (sentence) to Sn

Output: disease type d

Add disease flat mi method which is e.g. HIV, Cancer type return Boolean

Loop

Get Class methods m [] will be

Check mi is match d1. S1.word

New Disease return

End loop

Here conventional reflection pattern use for checking new disease type after adding its respective getter method into flat database.

The above task will be done by black box agent which will get all input from file system. Agent will get all disease type and Medline abstract.

5.7.2 Black Box Agent and Classification

Here, cyclic behavior means it will not die until user terminate application. Behavior method is used for business logic. The code in detail is in Appendix A. It explains the performative type, which indicates the message type such as INFORM_REF message.

In Appendix A, INFORM_REF indicates that message is coming and object of main document will be received. It is parent object and contains all predecessors' objects like document and Text i.e. sentences of Medline Abstract. Stats regarding document and clusters will be received. All clusters references also stored in document object. The data further populated into data grid in CreateHeader and set ColHeader () method.

5.7.3 Classification Results

Here clearly mention the domain and type of disease with proteins. The proteins are represented in column first and disease type is representing at column second “Disease”. In Figure 5.18, cancer and its type of diseases are represented. It shows the first cluster data.

the result were extracted by the process of mining. The classification results shown here indicate the process of k-mean with Jaccard approach. There are three clusters in total. The mention cluster shows the result of data about cancer. The result shown here is about ovarian cancer , prostate, bladder, colorectal , cervical cancer type. The all concern protein and its associated proteins from Medline abstract are about all cancer types.

Protein	Disease	
No Protein	ovarian	cancer
IMAC3	ovarian	cancer
7829	ovarian	cancer
6881	ovarian	cancer
A2780	ovarian	cancer
IMAC3	ovarian	cancer
7829	ovarian	cancer
6881	ovarian	cancer
001	prostate	cancer
001	prostate	cancer
001	prostate	cancer
001	prostate	cancer
5637/DR5	Bladder	Cancer
5	Bladder	Cancer
5637/DR5	Bladder	Cancer
5	Bladder	Cancer
mdr1	Bladder	Cancer
5637/DR5	Bladder	Cancer
DR5	Bladder	Cancer
56377	Bladder	Cancer
77	Bladder	Cancer
4	Bladder	Cancer
56377	Bladder	Cancer
R5	Bladder	Cancer
5637/DR5	Bladder	Cancer
5637/DR5	Bladder	Cancer
59	colorectal	cancers
30	colorectal	cancers
50	colorectal	cancers
70	colorectal	cancers
05	colorectal	cancers
05	colorectal	cancers
05	colorectal	cancers
p53	Cervical	Cancer
8±0	Cervical	Cancer
8±0	Cervical	Cancer
7±1	Cervical	Cancer
7±1	Cervical	Cancer
1±1	Cervical	Cancer
1±1	Cervical	Cancer
1±0	Cervical	Cancer
1±0	Cervical	Cancer
1±0	Cervical	Cancer

Figure 5.18: Classification Algorithm Results for Cancer Disease Cluster

In Figure 5.19, result indicates the cluster about HIV related protein. The “no protein” indicates the total recall versus precision. The result at end indicates the disease and disease type which is in this case is Extracellular HIV.

No Protein	prostate	cancer
No Protein	prostate	cancer
No Protein	prostate	cancer
No Protein	Bladder	Cancer
No Protein	Bladder	Cancer
No Protein	Bladder	Cancer
No Protein	colorectal	cancers
No Protein	colorectal	cancers
No Protein	colorectal	cancers
No Protein	colorectal	cancers
No Protein	colorectal	cancers
No Protein	Cervical	Cancer
No Protein	Cervical	Cancer
No Protein	Cervical	Cancer
No Protein	Cervical	Cancer
No Protein	Cervical	Cancer
No Protein	Cervical	Cancer
No Protein	Cervical	Cancer
No Protein	Cervical	Cancer
No Protein	Cervical	Cancer
No Protein	Cervical	Cancer
No Protein	ABreast	Cancer
No Protein	ABreast	Cancer
No Protein	ABreast	Cancer
No Protein	ABreast	Cancer
No Protein	ABreast	Cancer
No Protein	General	None
185	General	None
30	General	None
No Protein	in	Cancer
No Protein	in	Cancer
No Protein	in	Cancer
No Protein	in	Cancer
No Protein	in	Cancer
No Protein	in	Cancer
No Protein	in	Cancer
No Protein	in	Cancer
No Protein	in	Cancer
CD21	Extracellular	HIV
CD21	Extracellular	HIV
CD21	Extracellular	HIV
CD21	Extracellular	HIV
CD21	Extracellular	HIV

Figure 5.19: Classification Algorithm Results for HIV Disease Cluster

Figure 5.19 shows the third cluster which highlights the other disease data and wrong data which is not protein.

5.7.4 Classification Statistics

These clusters are shown in Figure 5.20 in different shades of pie chart graph. Red shows the general type of disease, blue showing the other disease like HIV. Green is showing

cancer type and its relevant diseases. The pie chart indicates the result of three type of clusters. The first cluster shows the percentage is 55% of protein which are related to cancer and its particular type. The other cluster indicates the percentage of 25% of HIV related proteins. The third clusters have precision from the total of 35%.

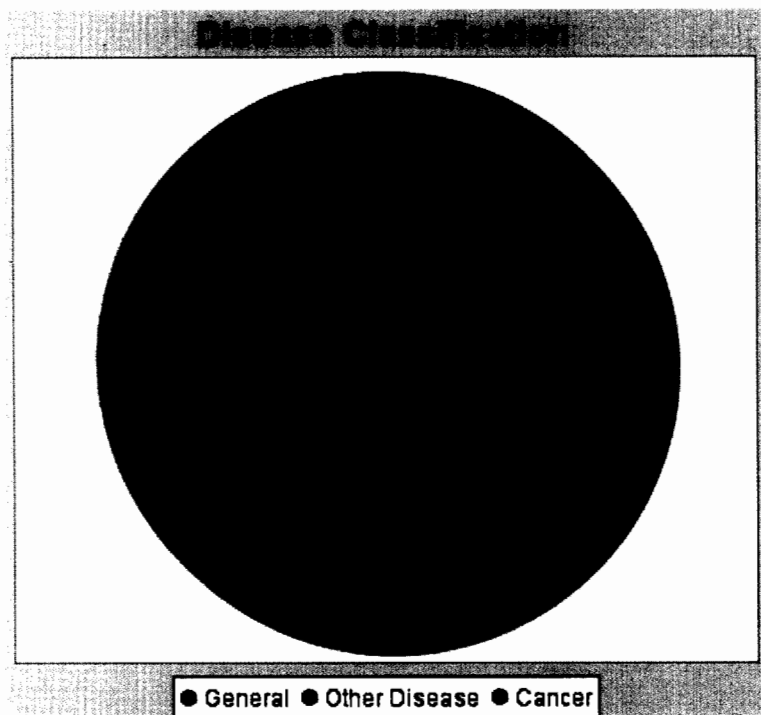


Figure 5.20 : Classification Pie Chart

5.7.5 Analysis of Results

We evaluated our final results and compare them with K-mean data mining approach and base paper approach of McCallum Unified Motivational approach. We experiment them using SPSS 9 and Abner Machine Learning tool. We evaluate results after getting Medline abstract from PubMed and Unbound Journal. We preprocess our results using technique of title analysis.

5.7.5.1 K-mean Data Mining vs. New Unified K-Mean

The k-mean algorithm shows the final output after analyzing 989 cases / sentences that were extracted from 30 Medline abstract using disease cancers, general disease. There were 6627 words in which have proteins in there sentences. But interesting thing in this

process is level of uncertainty which shown by the k-mean data mining procedure. In final output, few of proteins are not actually proteins but are cell lines. Due to less information about features of data it collapses. So final precession is less than New K-mean algorithm.

Number of Cases in each Cluster

Cluster	1	33.000
	2	437.000
Valid		470.000
Missing		220.000

K-Mean

Cluster	1	24.000
	2	258.000
Valid		282.000
Missing		210.000

New Unified K-mean

We use SPSS for K-mean algorithm analysis, the tools snapshots have shown in Appendix B. It shows that first we take sample which have numeric values and cell line as protein. Cell lines are e.g. C12 is cell line in abstract. The features about findings not measured because it is Euclidian distance based approach. We cluster the data based on distance. We did not measure feature about cell lines. We found almost 850 extractions which include numeric, cell lines and missing data. The level of uncertainty is so high as clearly shown in Figure 5.21.

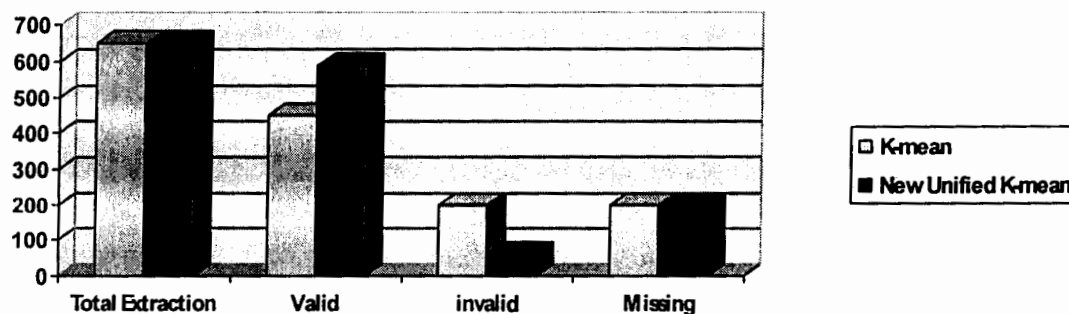


Figure 5.21: comparison between k-mean and New K-mean

The comparison summary shows the efficiency and effectiveness of results. We take 650

extractions out of which 450 seems valid extraction for k-mean data mining technique. But unfortunately this technique has invalid extraction from 450 valid extractions. There are 200 invalid extractions out of 450 extractions. There are 200 missing extraction we have found from total 850 extraction which we did not consider in results.

Figure 5.21 shows clear understanding of results that new unified k-mean algorithm applied over same sample. The results variation over valid protein is more than k-mean data mining. The results show the validity of unified approach and it also overcome uncertainty. Because it is more sensitive over extraction it extracts 588 entries and correctly attenuates the numeric entries and also cell lines. The cell lines have been attenuate or removed using features. There are 62 invalid entries which is cell lines in this sample.

Table 5.2: Comparison between k-mean vs. New Unified K-mean

	Cancer + General Disease Proteins	
	Precession	Recall
K-mean	70%	55%
New Unified K-mean	89%	91%

Table 5.2 shows the precision and recall comparison among k-mean and New Unified K-mean approach. Recall of new Unified K-mean is higher than K-mean. The main reason for this level of uncertainty is higher in k-mean data mining.

5.14.9.2 McCollum Unified Constraint Random Field vs. New Unified K-Mean

Using this approach we analyze 26819 words for 100 Medline abstract, during analysis we extract 2012 words from the 100 Medline abstract. The Medline abstracts are about cancer abrest/ovarian e.t.c and HIV related abstracts and General abstract about proteins. We analyze associations, protein functions, cell lines and stand alone proteins, synonyms related abstracts for these disease.

Cluster	1	429
	2	583
	3	512
Valid		1330
Missing		500

McCallum Unified Model

Cluster	1	437
	2	595
	3	562
Valid		1390
Missing		370

New Unified K-Mean

We use entity recognition tool which is Abner, this tool implement Machine learning technique of Constraint Random Field (CRF). See Appendix C for the tool annotation illustration. The constraint random field model is same model like one we found in MacCallum unification model of entity recognition [1].

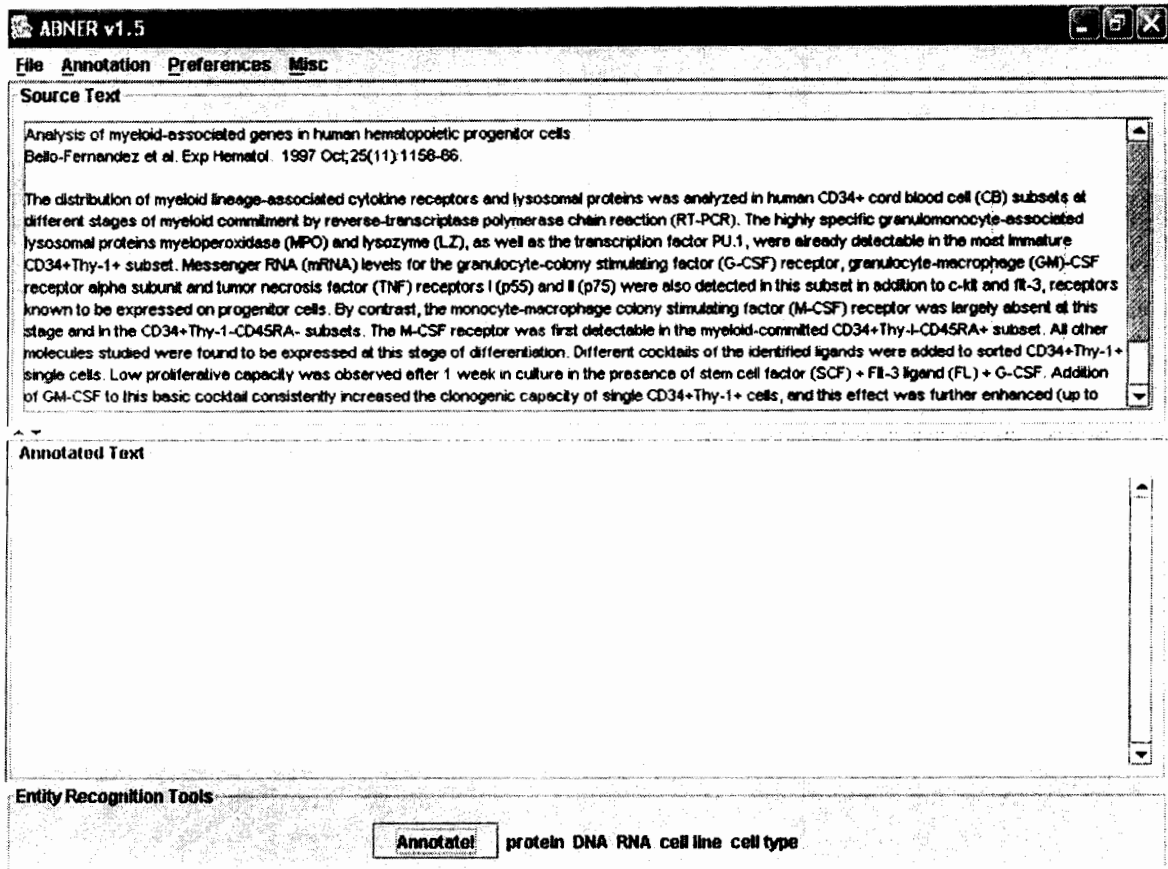


Figure 5.22: Medline Abstract Annotation by McCallum Model

We evaluated about every problem related to protein. We found standalone proteins, synonyms and associations. The constraint random tool annotates text by highlighting. But it did not unify two domains.

Figure 5.22 presents tool that first get text file as input source then it annotates the protein by highlighting in yellow text. We evaluated 100 Medline abstracts by this method. Finally we use our designed tool that extracts results from 100 Medline abstracts. We finally match input by searching findings from this highlighted text. After evaluation we found result very promising towards our new unified k-mean algorithm. In Figure 5.23, result analysis proof that the validity of Our Unified results is more than Constraint random Field (CRF).

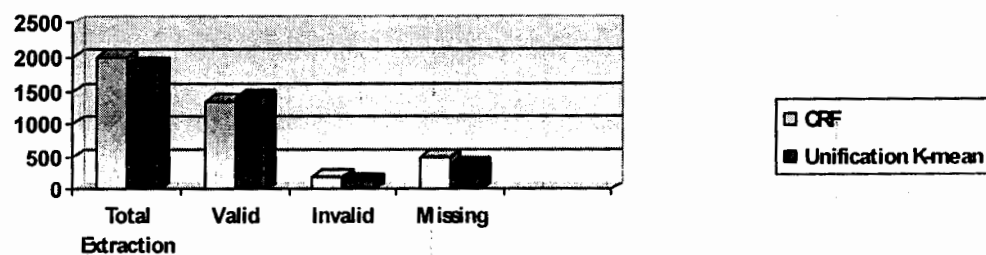


Figure 5.23: Comparison between McCallum Unified Model and New Unified K-mean

Figure 5.23 shows that classification results in case of Conditional Random Field are 2012. There were 500 missing entries we found in the result. These entries are extracted from Medline abstracts. Medline abstracts are downloaded from PubMed journal (one of biggest journal for Medline abstracts).

We found that few Medline abstract are about protein and its association. The McCallum Unified Model is not train for features about association rule because there is requirement of separate model and separate feature set for that. Hence, its valid results are 1330 which exclude the association of proteins. We evaluate same results with our approach which include those feature set regarding association. As in our objectives we clearly mention that there is requirement of single feature set for all type of protein. It extracts 1890 valid protein using New Unified K-mean Algorithm. The invalid results were 130 which include cell lines and cell expressions.

Table 5.3: Comparison between k-mean vs. New Unified K-mean

	Cancer+HIV + General Disease Proteins (Including Association)	
	Precession	Recall
McCallum Unified Model (CRF)	70%	85%
New K-mean	73%	90%

Table 5.3 shoes the precision and recall comparison between Constraint Random Field (CRF) and New Unified K-mean approach. Recall of new Unified K-mean is higher than Constraint Random Field. The main for this is exclusion of association of proteins from final extraction due to non availability of feature set.

5.8 Conclusion

There are few approaches which are related to information extraction and data mining, as we discuss in chapter 3. The feature identification was source from papers such as Automatic Extraction of Protein Interactions from Scientific Abstracts, a Web-resource for Exploring Protein Co-occurrences in MEDLINE Abstracts, Unsupervised gene/protein named entity normalization using automatically extracted dictionaries, A protein interaction extraction system and Automatic Extraction of Gene and Protein Synonyms from MEDLINE and Journal Articles. These papers were aid us to find the feature which were related to protein and its interaction.

The motivational paper which was base for unification and extraction helps to get the proper domain of interest. Automatic Extraction of Gene and Protein Synonyms are from MEDLINE and Journal Articles. The approach of unification was deriving from paper and rule of unification have five properties i.e. common procedure, common input and common output.

But some of approaches used either for information or information extraction or for KDD technique. A Mutually Beneficial Integration of Data Mining and Information Extraction but these are not satisfying the unification properly or producing results. A Note on the Unification of Information Extraction and Data Mining using Conditional-Probability, Relational Models.

The approaches which were discussed is about mining but for unification there are other approaches as well i.e. Mining Soft-Matching Rules from Textual Data, Un Yong Nahm and Raymond J. Mooney. These approaches are not impressive in terms of tight unification, so a approach is required to build feature matrix. Our approach checks similarity value using JACCORD approach. This is famous for binary variables. Finally database classification algorithm uses unification of data mining with information extraction using feature matrix. Feature matrix can also use with all other classification techniques.

5.9 Future Direction

In future the architecture will be mature with better unification model using adaptive learning of agent. This research unification algorithm support classification of three classes but it could be for n classes. The dynamic classification can be possible by hierarchical approaches of data mining. These points could be motivation from this research in future.

Appendix A Implementation Code

1. Black Box Agent and Classification

```

Protected void setup()
{
    sendMessage ("feature", list);

    addBehaviour (new CyclicBehaviour (this)
    {
        Public void action()
        {
            buildFeature= Recieve();

            if (msg! =null&&getPerfomative ()==13&& buildFeature!=null)
            {
                setPerformative(msg.DISCONFIRM);
                sendMessage("cluster",buildFeature);
            }

            Else if (msg! =null&&msg.INFORM_REF==getPerfomative())&&buildFeature!=null)
            {
                dialog1 = new ClusterDialog();
                List first = ((MainDocument) buildFeature).firstCluster();
                List second = ((MainDocument) buildFeature).SecondCluster();
                List third = ((MainDocument) buildFeature).ThridCluster();

                Vector col=new Vector();
                List total = ((MainDocument) buildFeature).getAlltext();

                int totaldoc=((MainDocument) buildFeature).getTotalDocuments();
                int totalrecords=((MainDocument) buildFeature).getTotalSentences();

                int totafirstcluster=((MainDocument) buildFeature).getFirst();
                int totasecondtcluster=((MainDocument) buildFeature).getSecond();
                int totathirdcluster=((MainDocument) buildFeature).getThrid();

                int firstoutcome=first.size();
                int secondoutcome=second.size();
                int thirdoutcome=third.size();

                col.add("Interactor");
                col.add("Disease");
            }
        }
    });
}

```

```

        col.add("Cancer");
        col.add("Interactee");

        Vector row=dialog1.createTableHeaders(first);
        dialog1.setColHeaders(row,col);
        final PieChart demo = new PieChart("Classification Results");

        Font f=new Font("Arial",0,10);
        JLabel l= RefineryUtilities.createJLabel("Mew",f);

        Panel p3=new Panel();
        p3.setBounds(50,300,400,400);
        demo.setBounds(100,100,400,400);
        DefaultPicDataset dataset=new DefaultPicDataset();

        demo.setDefault(buildFeature);
        demo.setVisible(true);
        dialog1.pack();
        dialog1.show();
    }
}

});
}

```

2. Jaccard Approach Implementation

```

public float JaccardApproach(int [] text1,int [] text2)
{
// jaccard approach applied here
float q=0f;
float s=0f;
float r=0f;
float t=0f;

for (int i=0;i<text1.length&& i<text2.length;i++)
{
if (text1[i] == text2[i] && text1[i] == 1 && text2[i] == 1) {
q++;
} else if (text1[i] != text2[i] && text1[i] == 0 && text2[i] == 1) {
s++;
} else if (text1[i] != text2[i] && text1[i] == 1 && text2[i] == 0) {
r++;
} else if (text1[i] != text2[i] && text1[i] == 0 && text2[i] == 0) {
t++;
}
}
}

```

```

    } else {
      ;
    }
  }

  return ((r+s)*0.5f)/((q+r+s)*0.5f);
}
}

```

Here r, s, q are binary comparison variable which use to measure feature similarity and dissimilarity.

```
float pointMean1 = JaccordApproach(getArray(pointText, mat.coloumn()),
getArray(meantxt1, mat.coloumn()));
```

```
float pointMean2 = JaccordApproach(getArray(pointText, mat.coloumn()),
getArray(meantxt2, mat.coloumn()));
```

```
float pointMean3 = JaccordApproach(getArray(pointText, mat.coloumn()),
getArray(meantxt3, mat.coloumn()));
```

3. Features Agent Implementation

```
public HashMap BuildFeature(ArrayList list)
{
  Iterator documentiter=list.iterator();
  while (documentiter.hasNext())
  {
    Document doc = (Document)documentiter.next();

    ArrayList txtlist=doc.getAllText();
    int j=0;
    Iterator textiter=txtlist.iterator();

    while (j<txtlist.size())
    {
      txt =(Text) txtlist.remove(j);

      StringTokenizer token=new StringTokenizer(txt.getSentence()," ,.")( "%+ -");

      while( token.hasMoreTokens())
      {
        String str =token.nextToken();

        txt.addWord(str);
      }
    }
  }
}

```

```

        RuleChecking(str, txt);
    }

    doc.getAllText().add(j,txt);
    j++;
}
doc.setAllTex(txtlist);
}
list=null;
return map;
}

```

4. Rule Checking by Feature Agent Ancestor

```

public boolean verb(String verb) {
    boolean flag = false;

    if ((verb.equals("bind to")) || (verb.equals("associate with")) ||
        verb.equals("interact with") || (verb.equals("characterise")) ||
        verb.equals("promote") || verb.equals("isolate") || verb.equals("suppress") ||
        verb.equals("modulate") || verb.equals("inhibit") || verb.equals("activate") ||
        verb.equals("regulate of") || verb.equals("regulate") || verb.equals("bind") ||
        verb.equals("associate") || verb.equals("interact") ||
        verb.equals("regulate through") || verb.equals("of regulate in") ||
        verb.equals("regulate of of") || verb.equals("regulate to") ||
        verb.equals("colocalize") || verb.equals("of to-regulate of") ||
        verb.equals("regulate through") || verb.equals("by regulate") ||
        verb.equals("regulate in") || verb.equals("to-regulate") ||
        verb.equals("regulator by") || verb.equals("regulators") ||
        verb.equals("regulate of in") || verb.equals("to-regulate in") ||
        verb.equals("of regulate of") || verb.equals("of regulate") ||
        verb.equals("regulates")) {
        flag = true;
    }
    return flag;
}

```

```
}

```

```
public boolean Synonym(String synonym , Text txt)

```

This method is use for synonym which will be analyzed in rule checking method. Below are scenarios which will run to show the different concepts of protein like interactor , interactee and synonyms.

```
public boolean Interactor(String InteractorChecker, Text txt)

```

This method is used to indicate that interactor is valid or not based on protein features. These feature are isDigit () , length () and hyphen.

```
public boolean Synonym(String synonym,Text txt)

```

5. Special Method and Feature Building

```
public boolean isNumeric(String numericCharacters)

```

```
{
    boolean numeric = false;
    int i=0;
    do
    {
        char c= numericCharacters.charAt(i);
        if (Character.isDigit(c)) {
            numeric = true;
        }
        else{return false;}
        i++;
    }while(i<numericCharacters.length());
    return numeric;
}
```


Next word will also indicating the Noun phrase as noun occurrences.

```
public boolean nextword(String next) {
    if ((next.equals("protein"))||
        (next.equals("Protein")) || (next.equals("cyclin")) ||
        (next.equals("CYCLIN")) || (next.equals("Proteins")) ||
        (next.equals("protiens")) || (next.equals("alpha")) ||
        (next.equals("beta"))) {
        return true;
    }
    return false;
}

public boolean measures(String measures) {
    if ((measures.equals("min")) || (measures.equals("kg")) ||
        (measures.equals("KG")) || (measures.equals("kda")) ||
        (measures.equals("KDA")) || (measures.equals("nl")) ||
        (measures.equals("NL"))) {
        return true;
    }
    return false;
}

public boolean compounds(String measures) {
    if ((measures.equals("sulphate")) || (measures.equals("urea")) ||
        (measures.equals("Sulphate")) || (measures.equals("Urea"))) {
        return true;
    }
    return false;
}
```

Appendix B K-Mean Analysis using SPSS

test-case1 [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1: V1 8402

	V1	V2	V3	var
1	8402	1	0	
2	BCRP/ABCG2	1	0	
3	K5	1	0	
4	K5	1	0	
5	K5	1	0	
6				
7	FLY1000	1	1	
8	ABCG2	1	1	
9				
10	ABCG2	1	1	
11	ABCG2	1	1	
12	BCRP/ABCG2	1	1	
13	BCRP/ABCG2	1	1	
14	GF120918	1	1	
15				
16	?	1	1	
17	17beta	1	1	
18	17beta	1	1	
19	?			
20	BCRP/ABCG2	1	1	
21	?			
22	Ko143	1	1	

K-Means Cluster Analysis

Variables: V1, V2, V3

Number of Clusters: 2

Method: Iterate and classify Classify only

Cluster Centers: Read initial

Write final

Iterate... Save... Options...

SPSS Processor is ready

start FinalDocumenta... Microsoft Of... test-case1 [Da... 11:47 PM

Appendix C Abner Constraint Random Field Tool (CRF)

ABNER v1.5

File Annotation Preferences Misc

Source Text

Analysis of myeloid-associated genes in human hematopoietic progenitor cells.
Bello-Fernandez et al. Exp Hematol. 1997 Oct;25(11):1158-66.

The distribution of myeloid lineage-associated cytokine receptors and lysosomal proteins was analyzed in human CD34+ cord blood cell (CB) subsets at different stages of myeloid commitment by reverse-transcriptase polymerase chain reaction (RT-PCR). The highly specific granulomonocyte-associated lysosomal proteins myeloperoxidase (MPO) and lysozyme (LZ), as well as the transcription factor PU.1, were already detectable in the most immature CD34+Thy-1+ subset. Messenger RNA (mRNA) levels for the granulocyte-colony stimulating factor (G-CSF) receptor, granulocyte-macrophage (GM)-CSF receptor alpha subunit and tumor necrosis factor (TNF) receptors I (p55) and II (p75) were also detected in this subset in addition to c-kit and flt-3, receptors known to be expressed on progenitor cells. By contrast, the monocyte-macrophage colony stimulating factor (M-CSF) receptor was largely absent at this stage and in the CD34+Thy-1-CD45RA- subsets. The M-CSF receptor was first detectable in the myeloid-committed CD34+Thy-1-CD45RA+ subset. All other molecules studied were found to be expressed at this stage of differentiation. Different cocktails of the identified ligands were added to sorted CD34+Thy-1+ single cells. Low proliferative capacity was observed after 1 week in culture in the presence of stem cell factor (SCF) + Flt-3 ligand (FL) + G-CSF. Addition of GM-CSF to this basic cocktail consistently increased the clonogenic capacity of single CD34+Thy-1+ cells, and this effect was further enhanced (up to

Annotated Text

Entity Recognition Tools

Annotates protein DNA RNA cell line cell type

Bibliography & References

- [1] Andrew McCallum and David Jensen, A Note on the Unification of Information Extraction and Data Mining using Conditional-Probability and Relational Models department of computer science, University of Massachusetts Amherst, 2003.
- [2] Razvan Bunescu, Edward M. Marcotte, Comparative Experiments on Learning Information extractions for Proteins and Their Interaction Department of Computer Science University of Texasm Austin USAs, Special Issue in the Journal Artificial Intelligence in Medline summarization and Information Extraction from Medical Documents, 2004.
- [3] Un Yong Nahm and Raymond J. Mooney, Mining Soft-Matching Rules from Textual Data Department of Computer Sciences University of Texas Austin, TX 78712-1188, Submitted to The Seventeenth International Joint Conference on Artificial Intelligence(IJCAI-01), 2001
- [4] Philip R. Cohen, Adam Cheyer, An Open Agent Architecture, Stanford University, 2004
- [5] Rayid Ghani , Rosie Jones , Dunja Mladeni , Kamal Nigam , Se'an Slattery, Data mining on symbolic knowledge extracted from the web, School of Computer Science _ Department for Intelligent Systems Carnegie Mellon University J. Stefan Institute Pittsburgh, PA 15213 USA Ljubljana, Slovenia, 2000
- [6] Andrew McCallum and Ben Wellner, Untangling text data mining. In *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.
- [7] Un Yong Nahm and Raymond J. Mooney. Toward conditional models of identity uncertainty with application to proper noun coreference. In *IJCAI Workshop on Information Integration on the Web*, 2003.
- [8] Fei Sha and Fernando Pereira. A mutually beneficial integration of data mining and information extraction. In *AAAI/IAAI*, pages 627–632, 2000.
- [9] David Heckerman, Christopher Meek, and Daphne Koller, Probabilistic Models for Relational Data, Microsoft Research and Stanford University Technical Report MSR-TR-2004-3, March 2004
- [10] B. Taskar, P. Abbeel, and D. Koller, Shallow parsing with conditional random fields. In *Proceedings of Human Language Technology, NAACL*, 2003.
- [11] Bhaskara Marthi, Brain Milch, Stuart Rusel. First-order Probabilistic Models for Information Extraction, University of California, 2004
- [12] Trausti Kristjansson, Aron Culotta, Paul Viola, Andrew McCallum, Interactive information Extraction with Conditional Random Fields, Microsoft Researah, 2003

- [13] David D. Palmer^{1,2} John D. Burger¹ Mari Ostendorf, Information Extraction from Broad Cast News Speech Data, The MITRE Corporation ²Boston University, 2000
- [14] Helena Ahonen Oskari Heinonen Mika Klemettinen A. Inkeri Verkamo, Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections Universität Tübingen University of Helsinki Wilhelm-Schickard-Institut für Informatik Department of Computer Science Sand 13, D-72076 Tübingen P.O. Box 26, FIN-00014 University of Helsinki Germany Finland, Contact: Manager, Copyrights and Permissions / IEEE Service Center / 445 Hoes Lane / P.O.Box 1331 / Piscataway, NJ 08855-1331, USA 1998
- [15] Lise Getoor Nir Friedman Daphne Kollar, Learning Probabilistic Relation Models, Computer Science Department, Stanford University, Stanford, CA. Division of Engineering and Applied Science Harvard University Cambridge, MA 02138. In Proc. IJCAI99, pages 1300–1309, Stockholm, Sweden, 1999
- [16] Georgios Sigletos, Georgios Paliouras and Vangelis Karkaletsis. Role identification from free text using hidden Markov Models, Software and Knowledge Engineering Laboratory Institute of Informatics and Telecommunications, N.C.S.R. “Demokritos”. 2002
- [17] James Thomas, David Milward, Christos Ouzunis, Automatic Extraction of Protein Interactions from Scientific Abstracts, Computational Genomics Group, The European Bioinformatics Institute, EMBL, University of Cambridge, 2000
- [18] A Web-resource for Exploring Protein Co-occurrences in MEDLINE Abstracts, Ardiane Genomics Tools for Systems Biology., 2001
- [19] Aaron M. Cohen, Unsupervised gene/protein named entity normalization using automatically extracted dictionaries, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA, *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature. Ontologies and Databases: Mining Biological Semantics*, pages 17–24, Detroit, June 2005.
- [20] Limsoon Wong, A protein interaction extraction system, Kent Ridge Digital Labs, Kent Ridge Digital Labs, Sigapor. 8 May, 2000
- [21] Hong Yu, Vasileios Hatzivassiloglou, Carol Friedman, Andrey Rzhetsky, W. John Wilbur, Automatic Extraction of Gene and Protein Synonyms from MEDLINE and Journal Articles. Dept. Medical Informatics, Columbia University, New York, NY 10032, USA, 2002
- [22] Dr, Sikandar Hiyat Khiyial, Sharjeel Imtiaz, Azmat Hussain. Using agents for unification of data mining and information extraction. International Islamic University Islamabad Pakistan. IEEE First International Conference. 2005
- [23] Jiawei Jan, Micheline Kamber, Data Mining concepts and techniques. A Book on Data Mining.

- [24] William W. Cohen, Haym Hirsh, Joins that Generalize: Text Classification Using WHIRL, AT&T Labs—Research 180 Park Avenue Florham Park NJ 07932 Department of Computer Science Rutgers University New Brunswick, NJ 08903, In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, August, 1998
- [25] Jason Rennie, Andrew Kachites McCallum, Automating the Construction of Internet Portals with Machine Learning, *Just Research and Carnegie Mellon University, Kluwer Academic Publishers. Printed in the Netherlands, 2000.*
- [26] Mikhail Bilenko and Raymond J. Mooney, Learning to Combine Trained Distance Metrics for Duplicate Detection in Databases, Department of Computer Sciences University of Texas at Austin Austin, TX 78712, February 22, 2002
- [27] Rakesh Agrawal, Tomasz Imielinski, Arun Swami, Mining Association Rules between Sets of Items in Large Databases IBM Almaden Research Center 650 Harry Road, San Jose, CA 95120, Proceeding of SIGMOD Conference Washington DC, 1993
- [28] Jennifer Neville and David Jensen, Iterative Classification in Relational Data, Knowledge Discovery Laboratory Department of Computer Science University of Massachusetts Amherst, MA 01003-4610, 1998
- [29] Scott Miller, Heidi Fox, Lance Ramshaw, and Ralph Weischedel. A Novel Use of Statistical Parsing to Extract Information from Text, BBN Technologies 70 Fawcett Street, Cambridge, MA 02138, 1997
- [30] Stefanie Brünninghaus and Kevin D. Ashley, Improving the Representation of Legal Case Texts with Information Extraction Methods Learning Research and Development Center, Intelligent Systems Program and School of Law University of Pittsburgh Pittsburgh, PA 15260, 2001, International Conference on Information Technology and Law, ACM, 2001
- [31] Andrew McCallum, Kamal Nigam, Lyle H. Ungar, Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching, WhizBang! Labs Research 4616 Henry Street Pittsburgh, PA USA, 2000
- [32] Jerry R. Hobbs, Douglas Appelt, FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text, published on Citeseer, Artificial Intelligence Center, SRI International Menlo Park California, 1996



1st International Conference on Information
and
Communication Technologies



Years of Excellence
1955 - 2005

Leadership and Ideas for Tomorrow

Using Agents for Unification of Information Extraction and Data Mining

Sharjeel Imtiaz and Azmat Hussain

Dr. Sikandar Hyat

Department of Computer Sciences,

International Islamic University, Islamabad (IIUI)

Sharjeel_ii@hotmail.com, Azmat_zain@hotmail.com

Abstract—Early work for unification of information extraction and data mining is motivational and problem stated work. This paper proposes a solution framework for unification using intelligent agents. A Relation manager agent extracted feature with cross feedback approach and also provide a Unified Undirected graphical handle. An RPM agent an approach to minimize loop back proposes pooling and model utilization with common parameter for both text and entity level abstractions.

I. INTRODUCTION

Previously there are many motivation works for unification of Information Extraction and KDD. The use of unified, relational, undirected graphical models for information extraction and data mining, in which extraction decisions and data-mining decisions are made in the same probabilistic “currency,” with a common inference procedure—each component thus being able to make up for the weaknesses of the other and therefore improving the performance of both [Andrew McCallum, David Jansen]. A unified model has been proposed for unification of information extraction and data mining, there are some of practical algorithm already proposed for unification purpose like existing text-mining systems discover rules that require exactly matching substrings; however, due to variability and diversity in natural-language data, some form of soft matching based on textual similarity is needed [1-Un Yong Nahm and Raymond J. Mooney] and has presented initial results on integrating IE and KDD that demonstrate both of these advantages[2-Un Yong Nahm and Raymond J. Mooney]. There are some approached that either learn rules for information extraction and KDD, it might be possible some approaches are based soft matching rule to prescribe working model instead of hard matching rules. Soft matching allows discovery

of additional interesting rules that more accurately capture certain relationships. Allowing the discovery of soft-matching rules can eliminate the need for certain types of tedious data cleaning prior to knowledge discovery. [Un Yong Nahm Raymond J. Mooney]. Soft matching and other integrated text mining algorithm covers prospect of uncertainty and invalid structure as well.

We are proposing a framework that work with underlying model to gain efficacy and efficiency using agent based architecture. An Open Agent Architecture [Philip R. Cohen Adam Cheyer] proposes multimodal artitecture for the purpose of identifying Rule specification, predicting behaviour and reviewing history feature. It is motivation from underlying artitecture. An harlov markov model are producing feature based on directed cyclic graph but use of unified, relational, undirected graphical models for information extraction and data mining, in which extraction decisions and data-mining decisions are made in the same probabilistic “currency,”

Proposes an undirected constraint model for over coming features and co-reference resolution with new model solution. Our approach is tested from the probabilistic component is a graphical augmentation of the relational model [Probabilistic Models for Relational Data by david hacker man,Daphan kollar].

II. PREVIOUS WORK

PRM model provide a framework for relational data mining. [Learning Probabilistic Relational Models. Lise Gooter, Nir Friedman] In this respect Bayesian models are inadequate to properly model properly aspect of complex relational domains. PRM is now development

stubs that attribute-based Bayesian network representation to incorporate a much richer structure. These models allow dependencies among other related entity but this model is unable to represent uncertainty and noise.

Markov model also represent a approach for information extraction due to acyclic reason the results does incur proper feature in order to cope with auto-correlation and constraints [Role identification from free text using hidden Markov Models, Georgios Sigletons].

In past at mid and last of nineteen there are some other approaches also working in this manner like [comparative experiments on learning information extraction for proteins and their interaction Razvan M.Marotte] lack good precision at bad recall too like Rapier, BWI and SVM a neural based approaches for named entity recognition and most of them incur low precision due to uncertainty and lower standard of information extraction technique. Recently probabilistic model have introduced by McCollum series of Conditional Random Field to overcome these deficiencies.

III. INFORMATION EXTRACTION THROUGH PROBABILISTIC MODELS

There are some of models are purposes over identity uncertainty; a model in order to acquire that two citation refer to same publication or whether two author paper written by same author problem. A model proposes for combined probability and constraint probability model to check all terms identity uncertainty and also check disambiguates after selecting only second one author title [First Order Probabilistic Models for Information Extraction, Bhaskara Marthi].

A learning algorithm which measure non inference entities and measure low distance between two phrases looked grammatically close to each other and also check when same distance it also checked weight position also. When values equal it does partitioning emergence, so to formulate by feature measurement [Toward Conditional Models of identity uncertainty with application to proper Noun Co reference, Andre McCullum, Ben Wellner].

But all above models does not have proper measuring of identity uncertainty such as one is

used to identify only all other attribute dependencies among common one but not Particularly handle a phrase, in this regard some of work is driven on phrases measurement through distance metrics but requires maximum likelihood for this purpose.

IV. DATA MINING THROUGH PROBABILISTIC MODELS

There are some of models such as Probabilistic models, which estimate self relationships, restricted relationships and probabilistic relationships. If all of these properties incorporate into one model then it is called facet of model such as DAPER model [Probabilistic Models for Relational Data, David Hackerman, Christopher Meek] a Microsoft research at Stanford university. Daper model expand directed acyclic graph and also represent self relations through model and provide a plate model for relationships. Daper is context free independent graphical model.

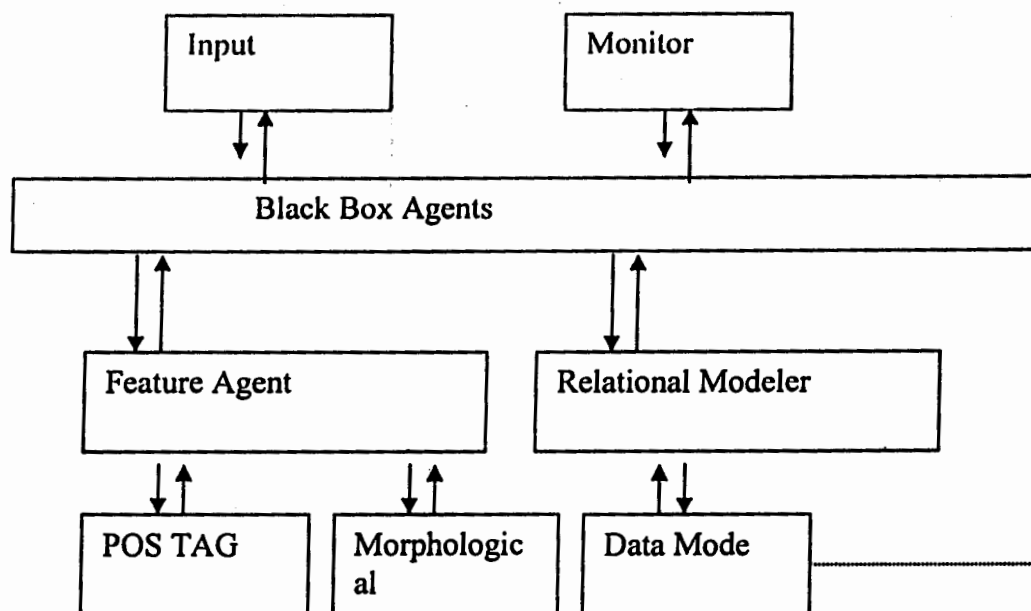
PRM models have many other forms such as Bayesian network [Learning Probabilistic Relation Model Lise Gooter, Daphne Kollar] in which parameter estimation occur in which structure is determine parents of attributes and maximum likelihood is the concept. It provides new ground to data mining after providing a rich class of relational models and statistical base.

V. AN AGENT BASED UNIFIED MODEL

Our approach is new convention to unification of previously developed approach and solution to problem. A unified Framework for both information extraction and data mining. It is because information extraction lacks constraints, co reference resolution and field extraction limited with noisy data.[Harlov] Data mining lacks features to cover uncertainty, limited in feature extraction and invalid structure. Unified framework purpose a model to produce multitude of label, Sequence of labels, binary resolution and cross reference.[Learning Probabilistic Relation Model Lise Gooter, Daphne Kollar].

A unified model with common input, common output effectively handle through agents and also common inference procedure for

both text and entity level abstraction. There are certain sorts of agents for undirected graphical model suitability.



Input type of agent provides data either hooked from entity attribute or either n-gram based tagged data to black box agent. Black box agent is special agent for purpose of cache structures, semantics, and syntax for the purpose of quick navigation from feature agent. Feature agent provide mode as further subsidiary agents POS Tag agent or morphological agent. POS tag agent work like hierarchal scheme if they handle tagged data feature then send back if not then send notification as rule specification to parent Feature agent. Then comes Morphological agents which handle particular request on Data type prefix and postfix features rules.

After binary features to black box it permits structure and passes feature agent query of model type to monitor agent. Monitor agent send rule specification for input data and parameters. Relation modeler select particular data model type after temporal variable reasoning about rules. It further passes feature and input training set to model which is either calculate estimated outcome to relation modeler as text mode or data mode. Inference as a results passed to black box which cache inference as it contain structures. Further monitor agent receives a notification of completion of inference procedure.

In order to persist changes it propagate to data source as shown in dotted line from data mode agent it also responsible for sending results directly to black box structure to improve performance.

This frame work improve inference and provides affinity or relationship matrix model and cross reference linear chain model as in relational modeler agent but separate the calculation of feature of attribute and relationship attribute by separating feature estimation and prediction of attribute.

VI. CONCLUSION

We have provided a frame work based on agents for unification purpose. It could handle feature estimation separately and prediction of attribute separately. In quality respect it provides Performance gain possible by caching and structure management at black box agents. Open architecture Hierarchal and communication architecture integration a motivation to previous architecture. It provides multiple modes of data handling through agents.

REFERENCES

- [1] A Note on the unification of information extraction and data mining using Conditional-Probability . Relational Models [Andrew McCallum, David Jensen]
- [2] Comparative Experiments on Learning Information Extractors for Protein and their Interactions [Razvan Bunescu].
- [3] Mining Soft-Matching Rules from Textual Data [Raymond J. Mooney]
- [4] Active Information Gathering in Infoscult [Marian Nodine, Jerry Fowler]
- [5] An Open Agent Architecture [Philip R. Cohen, Adam Cheyer]
- [6] Data mining on symbolic knowledge extracted from the web. [Hearst, 1999] Marti Hearst.
- [7] Untangling text data mining. In *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*, 1999. [McCallum and Wellner, 2003] Andrew McCallum and Ben Wellner.
- [8] Toward conditional models of identity uncertainty with application to proper noun coreference. In *IJCAI Workshop on Information Integration on the Web*, 2003. [Nahm and Mooney, 2000] Un Yong Nahm and Raymond J. Mooney.
- [9] A mutually beneficial integration of data mining and information extraction. In *AAAI/IAAI*, pages 627-632, 2000. [Sha and Pereira, 2003a] Fei Sha and Fernando Pereira.
- [10] Shallow parsing with conditional random fields. Technical Report CIS TR MS-CIS-02-35, University of Pennsylvania, 2003. [Sha and Pereira, 2003b] Fei Sha and Fernando Pereira.
- [11] Shallow parsing with conditional random fields. In *Proceedings of Human Language Technology, NAACL*, 2003. [Taskar et al., 2002] B. Taskar, P. Abbeel, and D. Koller.
- [12] Discriminative probabilistic models for relational data. In *Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI02)*, 2002.
- [13] First-order Probabilistic Models for Information Extraction, [Bhaskara Marthi, Brian Milch, Stuart Russell] University of California.
- [14] Interactive information Extraction with Conditional Random Fields, [Trausti Kristjansson, Aron Culotta, Paul Viola, Andrew McCallum], Microsoft Research.
- [15] Information Extraction from Broad Cast News Speech Data, [David D. Palmer, John D. Burger, Mari Ostendorf].
- [16] Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections [Helena Ahonen, Oskari Heinonen, Mika Klemettinen].
- [17] Learning Probabilistic Relation Models, [Lise Getoor]. Probabilistic Models for Relational Data, [David Hackerman].
- [18] Role identification from free text using hidden Markov Models, [Georgios Sigletos].

