AUTOMATED NURSING KNOWLEDGE CLASSIFICATION USING INDEXING

by

Sucharita Vijay Chinchanikar

A Thesis Submitted to the Faculty of

The College of Computer Science and Engineering

in Partial Fulfillment of the Requirements for the Degree of

Master of Science

Florida Atlantic University

Boca Raton, Florida

May 2009

ACKNOWLEDGEMENTS

It is a pleasure to thank the many people who made this thesis a success. I am indebted to my supervisor Dr. Shihong Huang for giving me this wonderful opportunity to work under her guidance throughout my Master's thesis. Her enthusiasm, inspiration and her great efforts to explain things clearly and in a simple way helped me to achieve my goals in this study.

Thanks to Dr. Abhijit Pandya and Dr. Sam Hsu for providing constant support and offering right direction. This work would not have been possible without them. I would also like to thank Dr. Fuhrt for his showing me the right direction and Jean Mangiaracina for her guidance through administrative hurdles.

Thanks to all my supervisors in Datacore Software Inc. and especially to Bettye Grant for all the understanding and support.

And last but not the least I would like to thank Dr. Marilyn Parker and all the wonderful people from nursing institute who helped me in understanding the nursing related part of my thesis. And of course to my family, thanks for believing in me.

ABSTRACT

Author: Sucharita Vijay Chinchanikar

Title: Automated nursing knowledge classification using indexing

Institution: Florida Atlantic University

Advisor: Dr. Shihong Huang

Degree: Master of Science

Year: 2009

Promoting healthcare and wellbeing requires the dedication of a multi-tiered health service delivery system, which is comprised of specialists, medical doctors and nurses. A holistic view to a patient care perspective involves emotional, mental and physical healthcare needs, in which caring is understood as the essence of nursing. Properly and efficiently capturing and managing nursing knowledge is essential to advocating health promotion and illness prevention. This thesis proposes a document-indexing framework for automating classification of nursing knowledge based on nursing theory and practice model. The documents defining the numerous categories in nursing care model are structured with the help of expert nurse practitioners and professionals. These documents are indexed and used as a benchmark for the process of automatic mapping of each expression in the assessment form of a patient to the corresponding category in the nursing theory model. As an illustration of the proposed

methodology, a prototype application is developed using the Latent Semantic Indexing (LSI) technique. The prototype application is tested in a nursing practice environment to validate the accuracy of the proposed algorithm. The simulation results are also compared with an application using Lucene indexing technique that internally uses modified vector space model for indexing. The result comparison showed that the LSI strategy gives 87.5% accurate results compared to the Lucene indexing technique that gives 80% accuracy. Both indexing methods maintain 100% consistency in the results.

To my husband Mr. Rohit Koppal

for his everlasting support and constant motivation.

AUTOMATED NURSING KNOWLEDGE CLASSIFICATION USING INDEXING

LIST OF FIGURES

Title **Page**

CHAPTER 1

INTRODUCTION

Embedding nursing language within informatics structures is essential to make the work of nurses visible, and articulate evidence about the quality and value of nursing in the care of patients, groups, and populations.

- Swan, Lang & McGinley (2004) [1].

1.1 Introduction

Providing people with a complete and responsive healthcare solution for the promotion of healthcare and wellbeing require the dedication of a multi-tiered healthcare service delivery system comprised of health professionals such as specialists, medical doctors, and nurses. Holistic approach to healthcare requires a comprehensive view of a patient care, including emotional, mental, and physical care needs. In addition to medical needs, friends, families and environment are aspects integral to the caring process, in which caring is understood as the essence of nursing.

The data collected by the nurse in assessment period is referred to as nursing knowledge, as it represents the knowledge of the nurse about the patient. As indicated by the number of nursing theories, each individual is different and hence should be treated differently. With the help of this nursing knowledge and the existing nursing theories, the nurse structures a holistic care plan that describes numerous characteristics of the patient and hence the type of care needed by the patient. This care plan helps the nurse to provide

the patient with the best care needed and help in his speedy recovery. Normally there are four critical elements of planning which include [2]:

- Establishing priorities
- Setting goals and developing expected outcomes (outcome identification)
- Planning nursing interventions
- Documenting

The nursing knowledge will guide the nurse through all these steps for planning. Nursing knowledge, which is formed during the interaction between the nurses and patients and their families, often is captured in natural language and in textual format. Proper capturing and managing this knowledge is essential for nurses to communicate among themselves, and with the rest of healthcare providers in order to provide a proper diagnosis and treatment plan. One of the steps is to map the textual assessment data to existing nursing theories to evaluate nursing practice. This Chapter describes the current approach to capture and manage nursing knowledge and issues faced by this approach.

1.2 Problem Statement

One of the methods used to analyze the nursing knowledge is use of printed paper forms. These forms outline categories of the nursing theory models along with the series of check boxes and empty spaces where nurse can manually enter the expressions from the assessment data and map it to the respective category.

While this method is the simplest to use, it has significant drawbacks. This method is both time consuming as well as error prone as it depends largely on an

individual who analyzes the data, and thus creates inconsistent classification results. The same data when analyzed by multiple nurses may produce multiple results. Also storing of this data in paper form is associated with all kind of storage issues. Some computer aided software engineering tools are available to semi-automate this process of analyzing the nursing knowledge data. These tool address the storage issues and to some extent, the time consumption issues.

1.3 Proposed Solution

Proper capturing and managing this knowledge is essential for nurses to communicate among themselves, and with the rest of healthcare providers in order to provide a proper diagnosis and treatment plan. One of important step in capturing nursing knowledge is to map the textual assessment data to a pre-defined existing nursing practice theory. This process is currently highly manual and time consuming. Also this process makes the results highly dependent on the individual nurse that analyzes the textual assessment data. This might make the results inconsistent when analyzed by multiple nurses. In order to make this process more efficient and consistent, we need to automate the classification of nursing knowledge with respect to the nursing practice model. This research focuses on automating the classification of nursing knowledge using an information retrieval technology known as Latent Semantic Indexing (LSI).

1.4 Thesis Contribution

The contributions of this thesis can be summarized as follows:

1. Defining a framework for automation of classification of nursing knowledge to make the classification more efficient and consistent. The documents are structured to include the words that best describe the categories in the nursing theory. This is achieved with the help of some expert nurse practitioners and nursing college professors.

2. Proposing a unique algorithm to classify the nursing knowledge automatically and accurately using the document indexing techniques. The proposed algorithm is implemented using Latent Semantic Indexing technique.

3. Validating the proposed framework in a real nursing case situation based on existing nursing theory. The simulation results are compared to measure the accuracy. The results were found to be 87.5% accurate.

1.5 Thesis Organization

Chapter 1 gives a general overview of the research by explaining what the nursing knowledge is and why it is so necessary to capture and manage it. It also describes the way it is captured and managed in today's life and the issues faced by the current approaches and the possible solution to address the issues.

The rest of the thesis is organized as follows:

Chapter 2 describes different categories in nursing theory and explains some of the existing nursing theories. It also describes some of the current representative technologies used to capture the nursing knowledge. It also points out some computer aided tools and technologies that are currently being used to capture or classify the nursing knowledge. The chapter also discusses the associated drawbacks of using these

technologies. This chapter also discusses three candidate technologies available in order to apply the solution: Semantic Web, Bayesian Classification and Latent Semantic Indexing.

Chapter 3 presents a framework for automatic classification using document indexing technique. The documents containing the respective category definitions along with the words that best describe the concepts and nursing language expressions are formatted by some expert nurse practitioners and professors. The documents are then indexed and used for the automatic classification.

Chapter 4 details the implementation of the proposed framework with the help of Lucene, an open source information retrieval library and exploiting Latent Semantic Indexing technique. The results are analyzed and are also compared to measure their accuracy.

Finally, Chapter 5 summarizes the research work. It also discusses implications of this research to the nursing communities and finally concludes with the identified limitations of this thesis and points to the future scope of study to improve these limitations.

CHAPTER 2

BACKGROUND AND RELATED WORK

To be master of any branch of knowledge, you must master those which lie next to it; and thus to know anything you must know all.

-Oliver Wendell Holmes Jr.

Nursing knowledge, which is formed during the interaction between the nurses and patients and their families, often is captured in natural language and in textual format. Proper capturing and managing this knowledge is essential for nurses to communicate among themselves, and with the rest of healthcare providers in order to provide a proper diagnosis and treatment plan. One of the steps is to map the textual assessment data to existing nursing theories to evaluate nursing practice. This section presents the background of some general approaches to nursing knowledge management, and describes representative methodologies for natural language classification.

2.1 The importance of capturing nursing language

Nursing is considered as an altruistic profession and the care given by a nurse to the patient is implied as holistic healthcare. A nurse considers the physical, emotional, social, economic, and spiritual needs of the patient [3]. She also considers the response a patient has to his or her illness, and also his ability to meet self-care needs. As a result,

the care provided by the nurse might compliment to that by a doctor. With an eye toward balancing quality and cost, healthcare planners are relying increasingly on nurse practitioners as the providers of choice for a range of front-line health services, such as primary and preventive care, managing chronic health conditions, and teaching older patients how to avoid injury and reduce the expenses of hospitalization and nursing home care. Mounting studies show that the quality of nurse practitioner care is equal to and at times better than, comparable care by physicians, and often lower cost [4].

Despite of this tremendous contribution of nurse to the quality of the healthcare, the work done by a nurse remains concealed. So the first aspect of capturing the nursing knowledge is to identify and analyze nurses' interventions in improving quality of healthcare. A standardized language is introduced to describe the care that is provided by nurses [5]. This standard is backed up by the various nursing theories.

Nursing documentation provides the basis for nurses to communicate with each other and with the rest of healthcare communities. So the second aspect of capturing the nursing knowledge is to structure a care plan for the patient to apply the knowledge of the existing nursing theories in order to provide better and holistic care to the patient.

2.2 Existing Nursing Theories

A theory in general could be defined as an idea that explains experience, interprets observations, describes relationships and projects outcomes [6].The nursing theories describe what is and what is not nursing. The definition of nursing theory as described in [7] states that "Nursing theory is a conceptualization of some aspect of

reality (invented or discovered) that pertains to nursing. The conceptualization is articulated for the purpose of describing, explaining, predicting, or prescribing nursing care problem statement." While [8] describes it as "Nursing theory is an inductively and/or deductively derived collage of coherent, creative, and focused nursing phenomena that frame, give meaning to, and help explain specific and selective aspects of nursing research and practice." The different type of nursing theories include: grand theory, middle-range theory and practice theory.

2.2.1 Grand theories

Grand theories have a very broad scope and provide some useful insights. Hence they are used for directing and predicting nursing in some particular situations but are generally not used for empirical testing.

Some examples of the theories that are considered to be at this level are as follows:

Leininger's theory of Culture Care Diversity and Universality

This theory describes the diverse and universal care factors that influence the health, well being and death of an individual. This theory states that, identifiable differences in caring values and behaviors between and among cultures lead to differences in the nursing care expectations of the care-seekers. The theorist holds that cultural care provides the broadest and most important means to study, explain, and predict nursing knowledge and concomitant nursing care practice. So the ultimate goal of this theory is to provide culture congruent nursing care practices [9].

Newman's Theory of Health as Expanding Consciousness

This theory was stimulated by concern for those for whom health as the absence of disease or disability is not possible. This theory is all about becoming more of oneself and helping patient to find greater meaning in life and connecting with other people in the world. The theory asserts that every person in every situation, no matter how disordered and hopeless it may seem, is part of the universal process of expanding consciousness [32].

Roger' Science of Unitary Human Beings

This theory assumes that human beings and their environment are infinite energy fields in continuous motion and produce unitary patterns. Rogers' three principles are the principles of resonancy (continuous change from lower to higher frequency), helicy (increasing diversity), and integrality (continuous process of the human and environmental fields) [33].

Some more examples of the grand theories are Orem's Self-Care Deficit Nursing Theory, Parse's Theory of Human Becoming.

2.2.2 Middle-range theories

Middle-range theories are narrower in scope than grand theories but broad enough to be useful for complex situations and appropriate for empirical testing. A wide range of situations and issues faced during nursing practice are handled by these theories.

Some examples of the theories that are considered to be mid-range theories are as follows:

Cristina Sieloff's Theory of Group Power within Organizations

This theory explains how to effectively manage the negative consequences due to changing health trends on the ability of an organization to achieve its goals. The goal of the theory is to achieve the “events that are valued, wanted and desired“ by utilizing the knowledge and skill of a group [10].

Katharine Kolcaba’s Theory of Comfort

This theory explains the intentional comfort actions that will help nurses to strengthen the patients and their families so that they can better engage in health seeking behaviors [11].

This theory was based on the work of earlier nurse theorists like Orlando (1961), Benner, Henderson, Nightingale, Watson (1979), Henderson, Paterson and non-nursing influences on Kolcaba's work included Murray (1938). The theory was developed using induction (from practice and experience), deduction (through logic), and from retroaction concepts (concepts from other theories) [12].

Ramona T. Mercer’s Theory of Maternal Role Attainment is also an example of the middle range theory.

2.2.3 Nursing practice theories

Nursing practice theories have the narrowest scope and are used within specific range of nursing situations. Nursing questions, actions and procedures are developed into nursing practice theories. Such theories are sometimes also called perspective theories [13, 14], micro-theories [15], situation-specific theories [7]. Nursing Practice theories are developed by quiet reflection on practice, remembering and noticing features of nursing

situations, attending to one's own feelings, reevaluating the experience and integrating the new knowing with other experience [16].

Parker/Barry Community Nurse Practice Model is an example of the nurse practice theory. This theory is used as the foundation of this research. The detail description of this theory will be discussed again in Chapter 3.

There are some computer aided software engineering tools available that can be used for qualitative analyses of data such as ATLAS.ti. These tools when used can address the storage issues and also the time consumption issue to some extent. This tool is discussed in next section.

2.3 Existing CASE tool support

ATLAS.ti is a software tool used for the qualitative analysis of large bodies of textual, graphical and audio/video data. An early prototype of ATLAS.ti was developed as part of the ATLAS Project (1989-1992) at the Technical University of Berlin [17]. It was later released as a commercial product by Thomas Muhr in 1993. It has multiple functions to administer, extract, compare and aggregate the meaningful data from a collection of data.

In order to use ATLAS.ti to classify nursing knowledge, the text file containing the patients assessment data is linked to the project in ATLAS.ti called Hermeneutic Unit (HU). This Hermeneutic Unit (HU) is the project and the documents, codes, memos, and other files associated with it. And the main workspace area is called Hermeneutic Unit (HU) Editor.

The Main Menu gives us the access to various components in HU which are the documents, expressions and codes. Then the document can be viewed in the primary document pane. Each expression in the file is analyzed by a nurse practitioner to map it to a related category from the nursing theory model. The data can then be queried, sorted or represented in a diagram.

This approach addresses the storage issues by storing the data electronically. It also addresses the time consumption issues to some extent as once mapped; the data can then be queried, sorted or analyzed using multiple electronic tools. But one major concern using this software for such purpose is preserving the consistency of the data. If the same data is analyzed by different people, the final outcome should remain unchanged. In order to achieve the required consistency, this category assignment needs to be automated. In order to automate the process of classification, we need to extract the information contained in each category of the nursing theory as well as the information in contained in each expression in the assessment data.

2.4 Related methodologies used for information retrieval

Among the several information retrieval technologies available today, the following are some technologies that are predominantly used for qualitative data analysis:

- Semantic Web Approach
- Bayesian Network Approach
- Latent Semantic Indexing

Each of these technologies is discussed in detail in the following sections.

2.4.1 Semantic Web Approach

Semantic Web is a framework that makes it possible to store the data in machine understandable format. This allows the machine to process the requests more efficiently. As the name suggests, it is a web of information which understands the semantics or meaning of each document contained in the web.

Figure shows the layers involved in semantic web framework. As seen in the Figure 2-1, Semantic web uses URI (Uniform Resource Identifier) to uniquely identify each resource. To make the most use of data, the documents are described using XML (eXtensible Markup Language). XML is the official recommendation of W3C (World Wide Web Consortium) that allows the use of self-descriptive tags to describe the data.

RDF (Resource Description Framework) describes the resources on the web in a machine understandable format. It is particularly intended to represent the metadata about the resources. It uses URI to identify the resources. RDF provides XML format (RDF/XML) to exchange information. It uses XML for modeling the information in the form of statements having triples that is Subject, Predicate and Object. While processing the documents, this data is broken down into these triples to extract meaningful information from it. Like HTML, this RDF/XML can be processed by a machine and using URI's, can link pieces of information across the Web.

Ontology is a formal representation of a set of concepts within a domain and the relationships between those concepts [18]. It is a shared vocabulary which can be used to model a domain that is, the types of objects and their properties and relations. So here ontology defines the relationship between the resources specified in the XML documents in order to facilitate the information retrieval.

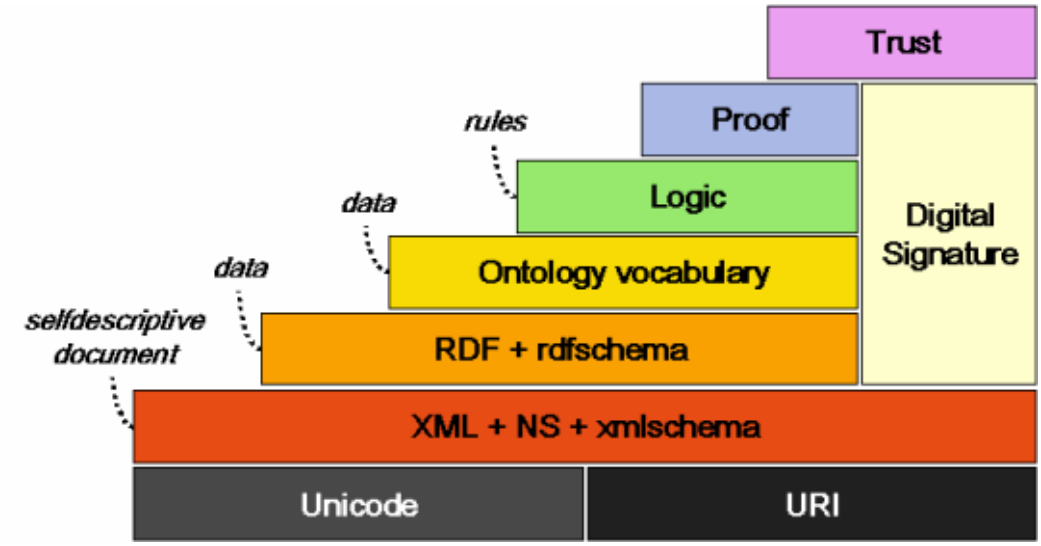

Figure 2-1 Semantic Web Layers

Logic layer is used to define the security rules on the data. For example, if some document mentions the sky is blue while the other mentions the sky is not blue, the machine should know which data to trust [19]. These rules are defines in the logic layer along with the use of Digital signatures. The proof layer verifies that the rules are followed while the final decision of whether the result is approved or not is made by the trust layer.

The domain ontology's represent the concepts in a very specific and often eclectic ways, hence they are often incompatible. So as the system expands the ontology's need to be merged in a more general representation. At present, merging the ontology is highly manual, time consuming and expensive process [18]. Also in order to incorporate this technology, we need to express the nursing knowledge as well as the nursing theories in

XML format. But as the nursing knowledge is written in plain natural language text, we need to seek a technology which can process the data in the form of natural language text.

2.4.2 Naïve Bayes' Text Classification

Bayesian network is a supervised learning method. It is a graphical representation that considers probabilistic relationship for variables of interest [20]. Bayesian network classifier obtains network structure and conditional probability table by learning training data set, and then implements the classification by using Bayesian theorem to compute posterior probability [21]. Naïve Bayes' classifier is a special type of Bayesian classifier. It is based on Bayes' theorem of probability

$$p(v/d) = \frac{p(v)\,p(d/v)}{p(d)} \tag{1}$$

Here $p(d/v)$ is the conditional probability of d given v. $p(v)$ and $p(d)$ represents probability of the terms and probability of documents in the whole collection.

If we write this equation in plain English, it would appear as follows:

$$posterior = \frac{prior \times likelihood}{evidence} \tag{2}$$

For example to classify the expression E as the document D with the highest posterior probability:

$$p(D_i/E) = \frac{p(D_i)\,p(E/D)}{p(E)} \tag{3}$$

The document can be represented according to Bernoulli document model as the binary vector whose elements are indicated by the presence or absence of the word inside

the document. Then $p(E/D)$ is the probability of expression E occurring inside the document D. $p(D_i)$ is the probability of occurrence of the document in the whole classification training examples. $p(E)$ is the normalized probability of occurrence of each expression in the whole classification training examples.

One common rule is to pick the hypothesis that is most probable; this is known as the maximum a posteriori (MAP) decision rule. The corresponding classifier is the function classify defined as follows:

$$\text{classify}(f_1, \ldots, f_n) = \text{argmax}_c \, p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c). \quad (4)$$

This technology is widely used for qualitative analyses of data but its accuracy depends on the amount of data used as a reference to classify. As the amount of reference data increases, the classification is more accurate. But if there is no sufficient reference data provided, then this could yield inaccurate results.

2.4.3 Latent Semantic Indexing

Latent Semantic Indexing (LSI) uses statistically derived conceptual indices to match the query and retrieve the information from a set of documents. A truncated Singular Vector Decomposition (SVD) is calculated to predict the structure of word usage in the document. It is also called Latent Semantic Analysis.

In the beginning, the LSI algorithm purges all the extraneous words from the document or title[22]. Then using the 'Porter Stemming' algorithm, the common endings

of the words are eliminated. This helps to eliminate the noise generating words and characters that hinders the decision making task [23]. Then a matrix is created with rows representing the terms in the document and the columns representing the documents. This matrix is called term-document matrix. Each value in the term-document matrix is a weighted frequency of the term in the corresponding document. This term-document matrix is reduced to a compressed matrix by applying SVD technique.

For example, if after eliminating stop words and applying stemming algorithm there are m distinct terms in n documents, the $(m \times n)$ term document matrix X is formed where each term x_{ij} in the matrix is weight of the term. Then the SVD of the matrix is given by,

$$X = U \sum V^T \quad \textbf{(5)}$$

Where U and V are matrices of the left and right singular vectors. $\sum$ is the diagonal matrix of singular values [24]. Here each column in the transpose of matrix V represents the each document in the database.

Similarly, query string is converted into a query Vector. Then the semantic closeness between the query string and each document is calculated by cosine similarity between the two vectors. The cosine similarity between the query vector q and Document vector d_j is given by,

$$Sim(q, d_j) = \frac{(q \bullet d_j)}{|q| \times |d_j|} \quad (6)$$

Here, $(q \bullet d_j)$ is the dot product of query vector and document vector. $|q|$ is the Euclidean norm of the query vector and $|d_j|$ is the Euclidean norm of the document

vector. The document having the highest cosine similarity is considered as semantically close document.

This technology can produce good and reliable results even with the minimal amount of data available. Hence we will use this technology to automate the classification of the nursing knowledge.

2.5 Summary

This chapter discusses the importance of capturing and managing nursing knowledge. It also describes the different types of nursing theories defined. Then it explains the computer aided tools that are currently available in the market which provide a means to semi-automate the process. It also discusses some other information retrieval technologies that could possibly automate the whole process of capturing the nursing knowledge in order to prepare a good and consistent nursing care plan document. It discusses the advantages and disadvantages of using these technologies for automating the classification of nursing knowledge.

CHAPTER 3

A FRAMEWORK FOR AUTOMATED CLASSIFICATION OF NURSING LANGUAGE

Four steps to achievement: plan purposefully, prepare prayerfully, proceed positively, pursue persistently.

-William A. Ward.

3.1 Introduction

This section presents a framework that automatically classifies the nursing data that is in the textual assessment form and mapping it to the corresponding category in the nursing knowledge model based on nursing theories. The approach focuses on increasing the accuracy and consistency of the classification of nursing knowledge. The architecture of this framework is based on the canonical activities of reverse engineering [25].

The accuracy of the classification largely depends on the structure of the documents that are used as baseline for indexing. We have structured the baseline documents such that each document represents a category in the nursing practice model. So we need to make sure that all the terms that describe the category are included in the corresponding document.

We will use document indexing technique to automate the classification of the assessment data based on the nursing model. We then index the documents and calculate the relevancy ranking of the documents with regards to each sentence in the assessment data. The sentences are then mapped to the documents (representing the categories) having highest relevancy ranking.

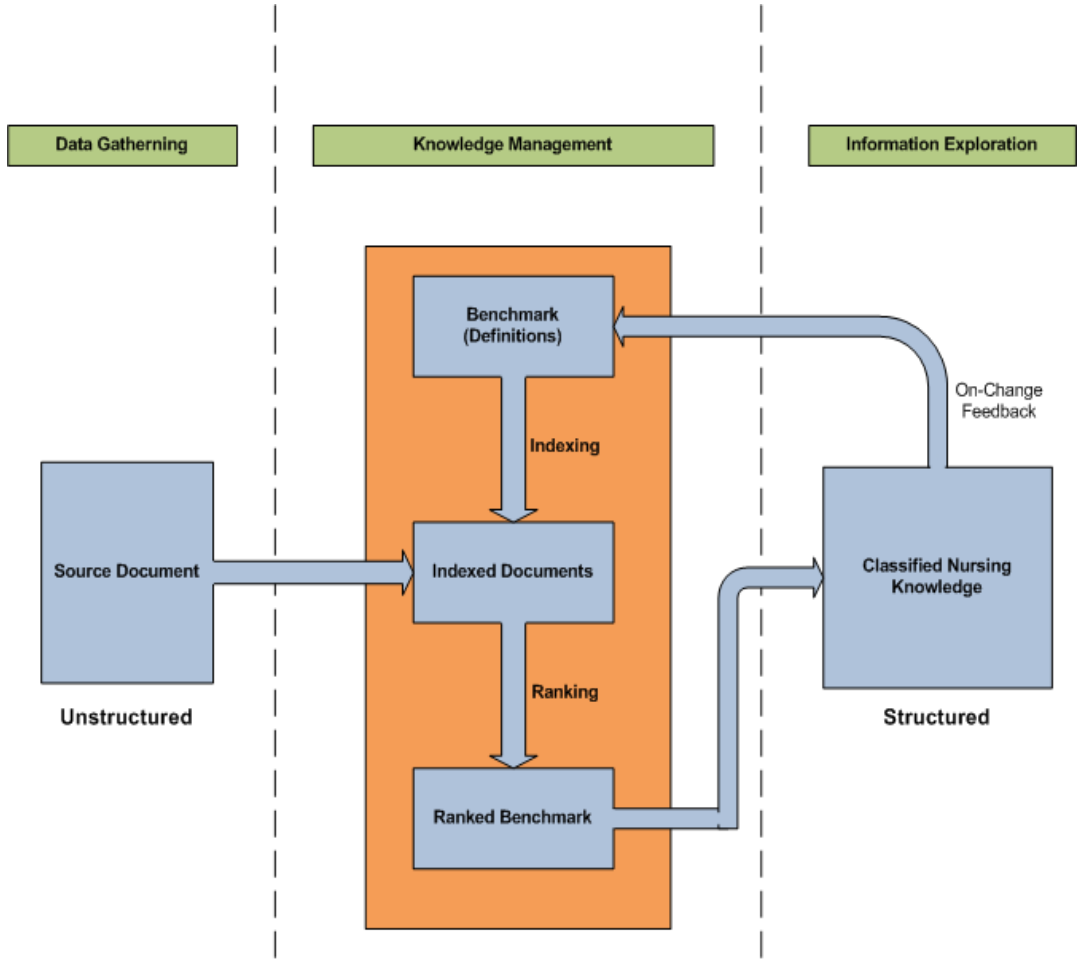

Figure 3-1 Framework for Automated Classification of Nursing Knowledge

3.2 Architecture

Figure 3-1 illustrates the framework for automated classification of nursing knowledge using document indexing technique. As shown in the Figure 3-1, the overall process can be divided into three logical steps: Data gathering, knowledge management and information exploration. The following sections describe these three steps in detail.

3.2.1 Data Gathering

Gathering data from the patient is the first and most essential step. The raw data gathered is the narrative written in natural language by nurse. This data is collected by nurse during the patient's assessment period. This data is called nursing knowledge as it represents the nurse's interpretation of the patient's condition, his history and his concerns. This data is crucial for nurses to understand the patient's needs and concerns. Based on this data, nurse can plan the nursing needs for the patient and document it. This raw data is the unstructured data and should be put into a representation that will facilitate efficient and easy storage and retrieval. Hence we will structure or classify the data based on a pre-defined nursing model. The nursing model captures the significant aspects of the nursing instrumental values. These various aspects will help the nurse to plan for the steps that she needs to take in order to provide the best care to the patient.

3.2.2 Knowledge Management

The nursing model is comprised of number of aspects. For example, as shown in Figure 3-2, Parker/Barry Community Nurse Practice Model is comprised of four main concepts:

- Caring
- Wholeness
- Connections
- Respect

These concepts are also known as the nursing instrumental values as they represent the four most important values of nursing practice.

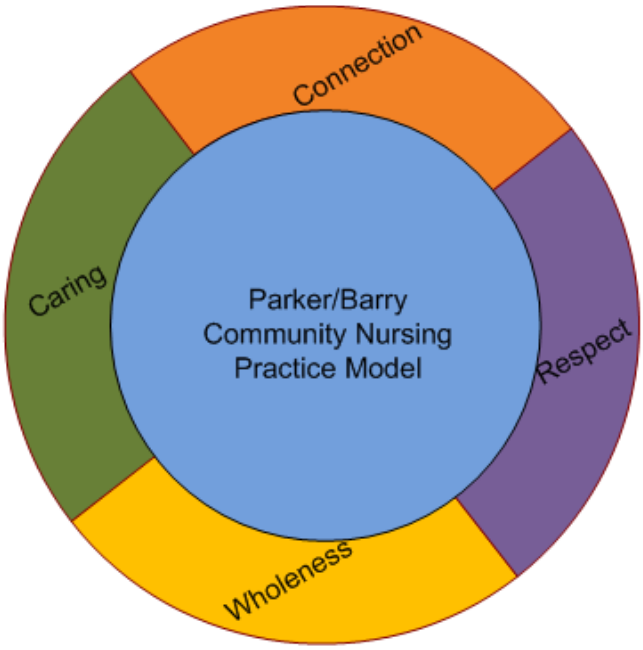

Figure 3-2 Parker/Barry Community Nursing Practice Model

In order to classify the nursing knowledge into these categories, we need to collect all the information pertaining to each category into a separate document. The

information could be a thorough definition of the category, all the terms that best describe the category along with the example expressions. Figure 3-3 shows the example expressions in "Respect" category in Barry/Parker Community Nurse Practice Model. Here some of the words that best describe respect category are: Honoring autonomy, listening, being courteous and respectful and preserving dignity. This collection of document will then act as a benchmark to help the further automation. The example expressions that represent these concepts are:

- We found a quiet place and provided time for planning – Honoring autonomy.
- I listened and heard her as I tried to learn about her life – Listening.
- I was respectful – Being courteous and respectful.
- I tried to keep him clean and dry – Preserving dignity.

All such best describing words along with the example expressions are included in the document representing the concerned category.

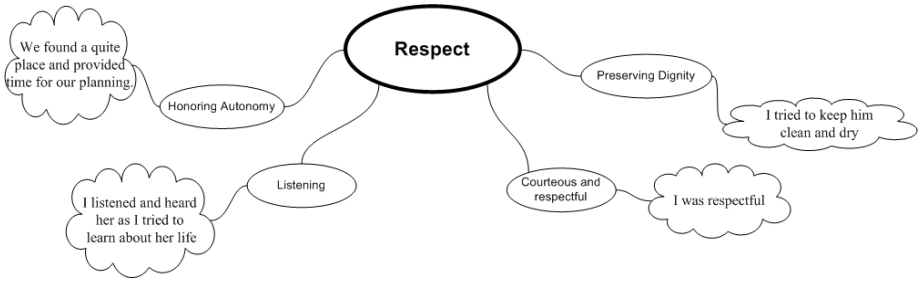

Figure 3-3 Elaboration of the Respect Category

These benchmark documents are then indexed using Latent Semantic Indexing technique which assigns indices based on the semantics of the document. As the nursing knowledge is expressed in natural text, different terms could be used to express the same

concept (synonyms). But as LSI technique can relate the terms co-occurring inside the documents, by using LSI technique we can ensure the correct classification the data even if the exact matching term is not present in the benchmark document.

Let us take an example to explain co-occurrence of terms. If two terms love and care co-occur in documents d2, they are clustered by LSI algorithm. Hence, if in some document d1, if the term love appears, all the documents containing word care are also included in the result. Figure 3-4 shows an illustration of this example.

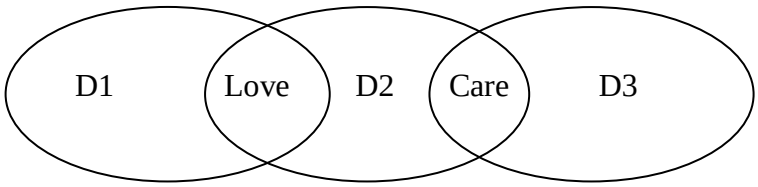

Figure 3-4 Co-occurrence of Terms Inside the Documents

Once the benchmark documents are indexed, the documents are ranked with respect to each sentence in the source document. The documents that are semantically close to the source sentence in the source document are ranked higher. So each sentence in the source document gets mapped to the concept represented by the highest ranked benchmark document.

3.2.3 Information Exploration

While data gathering is important to initiate the whole process and knowledge management is important to structure the data into a conceptual model, information

exploration is essential for the nurse to make the efficient use of the structured data. The tree structure representing the structured layout of the nursing knowledge will be presented to the nurse. This makes it easier for her to navigate through the data. Also as the nursing knowledge is sorted and grouped into the different categories in the nursing model, it will facilitate the clear understanding and quick planning. If the nurse feels that any of the sentences is incorrectly mapped, she can remap the sentence to the correct category. This in turn will trigger the feedback process where the corresponding sentence will be added to the document representing the correct category. This will ensure the correct classification of the similar sentences in future. This tree structure helps the nurse to get a single shot view of the patient's history and hence to generate a more efficient care plan for the patient.

3.3 Summary

In this chapter we discussed the framework that we proposed for the automated classification of the nursing knowledge using Latent Semantic Indexing technique. The overall process is divided into three phases: Data gathering, knowledge management and information exploration. The nursing knowledge is collected in the data gathering phase. In knowledge management phase, the information about the nursing model is collected in the benchmark documents. The benchmark documents are then indexed and ranked with respect to the source document. The resulting output is the structured layout of the nursing knowledge in the information exploration phase.

CHAPTER 4

CASE STUDY

For the things we have to learn before we can do them, we learn by doing them.

- *Aristotle.*
-

4.1 Data collection for simulation

To validate the proposed framework, the automated classification of nursing knowledge is simulated using Latent Semantic Indexing technique. The benchmark model used is "Parker/Barry community nurse practice model". The following sections describe the case study and discuss the simulation results and analysis.

4.1.1 Data collection of nursing knowledge

To test the accuracy of the simulation against diverse possible scenarios, we have assembled the test data collected and classified by numerous skilled nurse practitioners. This test data consists of 20 expressions of each category of "Parker/Barry Community Nursing Practice Model". Each of these expressions is different from the example expression given in the benchmark documents.

4.1.2 Data collection of benchmark documents

Parker/Barry Community Nurse Practice Model is comprised of four main concepts: Caring, Wholeness, Connections and Respect. These concepts are also known as the nursing instrumental values. The benchmark documents are arranged such that each document represents a category in Parker/Barry community nurse practice model. Each document is well structured to contain a thorough definition of the category along with the terms that best describe the category and the example expressions for the respective category. These documents will serve as the benchmark to classify the nursing knowledge.

4.2 Knowledge management

This part is the crux of the automation as it comprises the logic to automate the classification. Figure 4-1 shows the flowchart of our framework using Latent Semantic Indexing technique. Following are the definitions of some terms used in the flowcharts.

- Benchmark documents – The word documents which act as the reference and contain the definition for each category in the Nursing Model along with the example expressions and synonyms of the terms defining the category.
- Stop Words – The most commonly used English words which do not give any information about the documents. Example words: a, the, he, she, it.
- Stemming Algorithm – This algorithm reduces the inflected or derived words to their stem, base or root form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem,

even if this stem is not in itself a valid root [26]. For Example, the words like 'Nursing', 'Nurse' will be reduced to 'Nurs' and words like 'Happiness', 'Happy', 'Happily' will be reduced to 'Happi' after applying stemming algorithm.

- Term-document matrix – The $term \times document$ matrix with each universal term as its row and each benchmark document as its column where the value of each matrix element is (term frequency in respective document) / (overall term frequency).

  To explain this, let us take an example. Let the collection contain three documents. If the term 'Care' appears two times in the first document and once in document-2 and never in document-3, the row for the term care in the term-document matrix will be

$$row = \{2/3 \quad 1/3 \quad 0\}$$

- Universal terms – A list of all the unique terms in the benchmark documents.
- Expanded query – The collection of the terms obtained by collecting synonyms of each unique term in the query expression. The synonyms are obtained from the wordnet dictionary.
- Query vector – A binary vector prepared by comparing each term in the expanded query to each term in the universal terms list.
- SVD – Singular Vector Decomposition (SVD) is a process which decomposes a matrix into three different matrices.

Let us start with the creation of the term document matrix. The stop words are eliminated from the benchmark documents. This process will purge all the words that do not carry any information. Then the stemming algorithm is applied to the document to obtain a list of non-repeating unique terms in the collection of document. This collection of terms is designated as the universal terms.

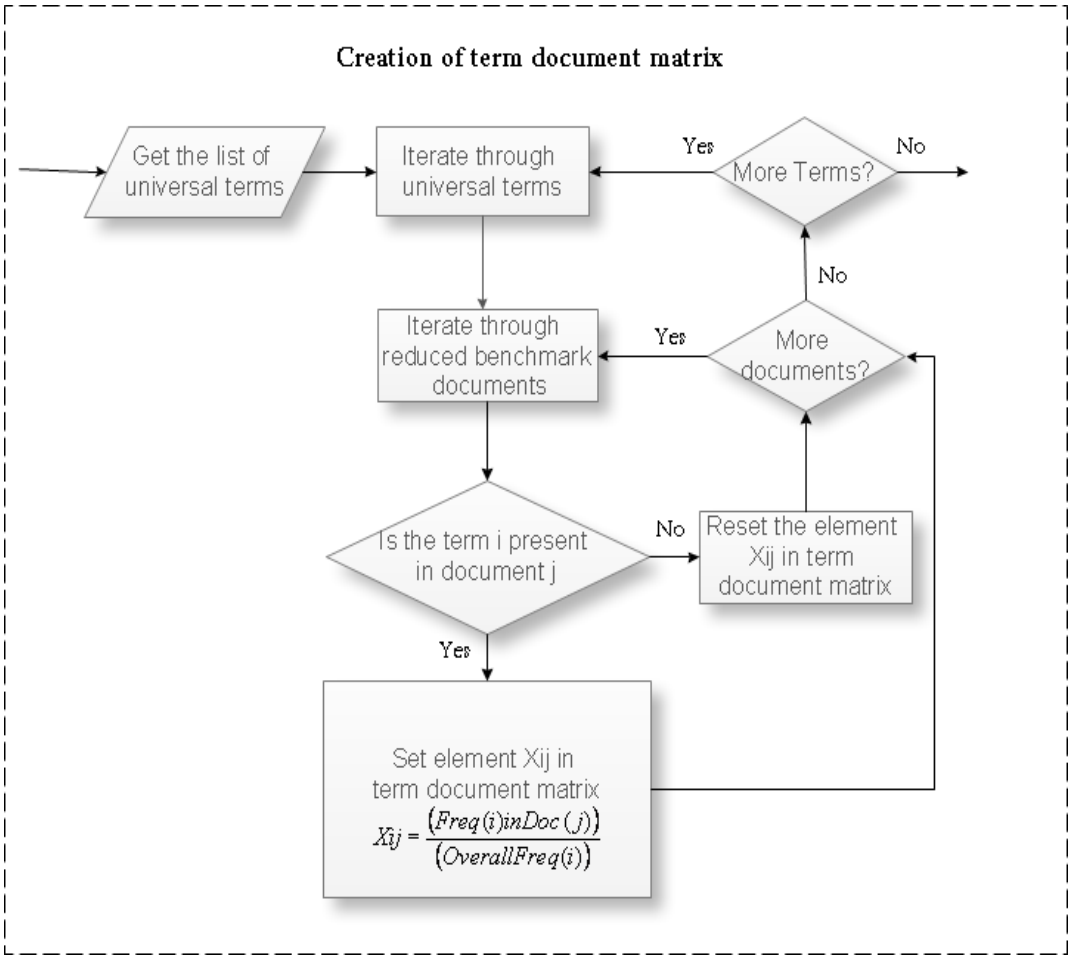

Figure 4-1 Creation of Term-document Matrix

Weight of each term is calculated by iterating through the universal terms and the term-document matrix is prepared. Figure 4-1 above explains this process of creation of term

document matrix in detail. After the completion of this process, we will have a matrix similar to the matrix shown in Figure 4-2, representing our benchmark data where the rows of matrix represent the unique terms inside the benchmark documents whereas the columns represent the benchmark documents.

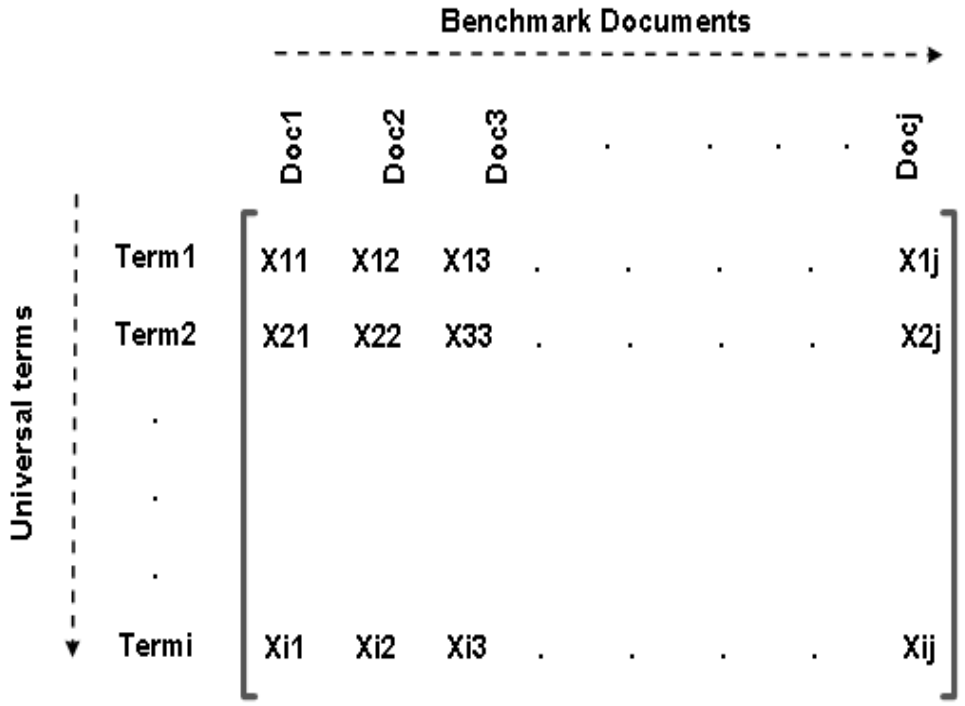

Figure 4-2 Term-document Matrix

Once the term document matrix is created, we will decompose the term-document matrix into three different matrices using singular vector decomposition (SVD) method. So when we decompose the term-document matrix with i rows and j columns, we will get

a matrix S, a diagonal matrix with same dimensions as of the term-document matrix which is i rows and j columns, and two unitary matrices U and V such that,

$$\text{Term-document matrix} = U*S*V \tag{7}$$

U will be a square matrix with i rows and i columns. So U will represent all the unique terms inside the benchmark database. Also the dimensions of V will be j rows and j columns. So matrix V will represent the documents in our benchmark collection.

Now we will create our query matrix using the information in the S and U matrices so that we can compare the similarity between the query matrix and information about every document in the V matrix.

To prepare the query matrix Q, we will first generate a query vector q by

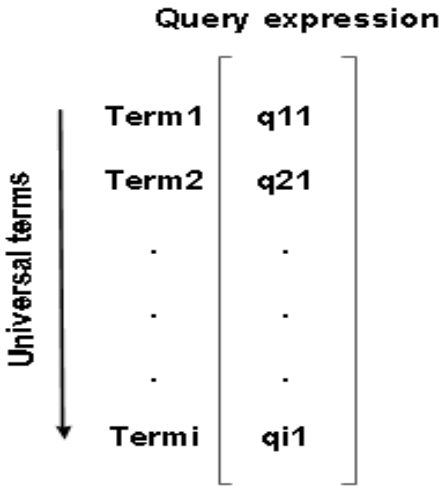

Figure 4-3 Query Vector

matching our query expression which is the nursing knowledge, to the universal terms list. If the term is present in the query expression, we will set the corresponding element in the query vector otherwise we reset the element.

Figure 4-3 shows the diagrammatic representation of the query vector. The query vector generated will be single column matrix with i rows. The query matrix Q is then generated from query vector q using the following equation.

$$Q = q^T \times S^{-1} \times U \tag{8}$$

Where S and U are the matrices generated from the SVD processing of the term-document matrix and q^T is the transpose of the query vector q.

The query matrix obtained will be a single column matrix with the number of columns same as the number of columns in the matrix V. This query matrix represents the co-ordinates of the query expression in the semantic space. Similarly, each column in

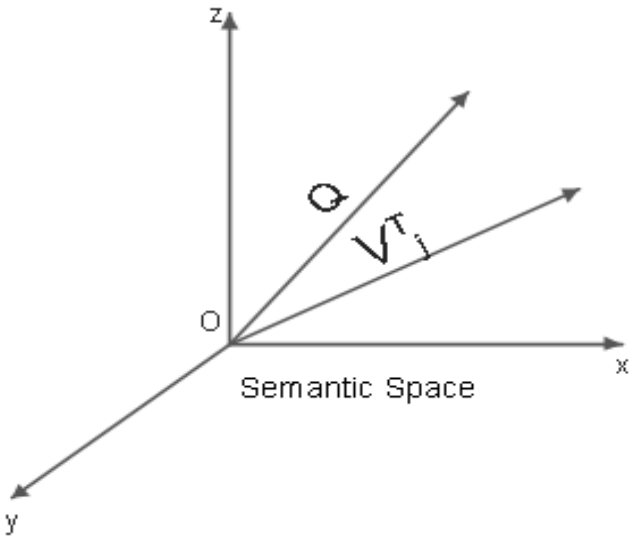

Figure 4-4 Query Vector and Document Vector in Semantic Space

the transpose of matrix V represents the co-ordinates of each benchmark document in the semantic space. Figure 4-4 shows the diagrammatic representation of query vector and document vector in the semantic space.

Now by calculating the cosine similarity between the two vectors, we can find the benchmark document that is most similar to the query expression. The cosine similarity between the query matrix Q and document matrix $V^T{}_j$ is given by the following equation:

$$Similarity(Q,V^T{}_j) = \frac{(Q \bullet V^T j)}{|Q| \times |V^T j|} \tag{9}$$

Here, $(Q \bullet V^T j)$ is the dot product of the two vectors and $|Q|$, $|V^T j|$ are the Frobenius Norms of Q and $V^T{}_j$ also called Euclidean Norms which give normalized unit vectors in case of the single column matrices.

After calculating the cosine similarities between the query matrix and each document matrix, we will have j different ranks for j benchmark documents against each query expression. We call this process as ranking the documents. The document with the highest rank is most similar to the query expression. So we assign the query expression to the category represented by that benchmark document.

Figure 4-5 shows the combined flowchart of all the processes we just discussed. Now let us see the step by step implementation of the processes just discussed.

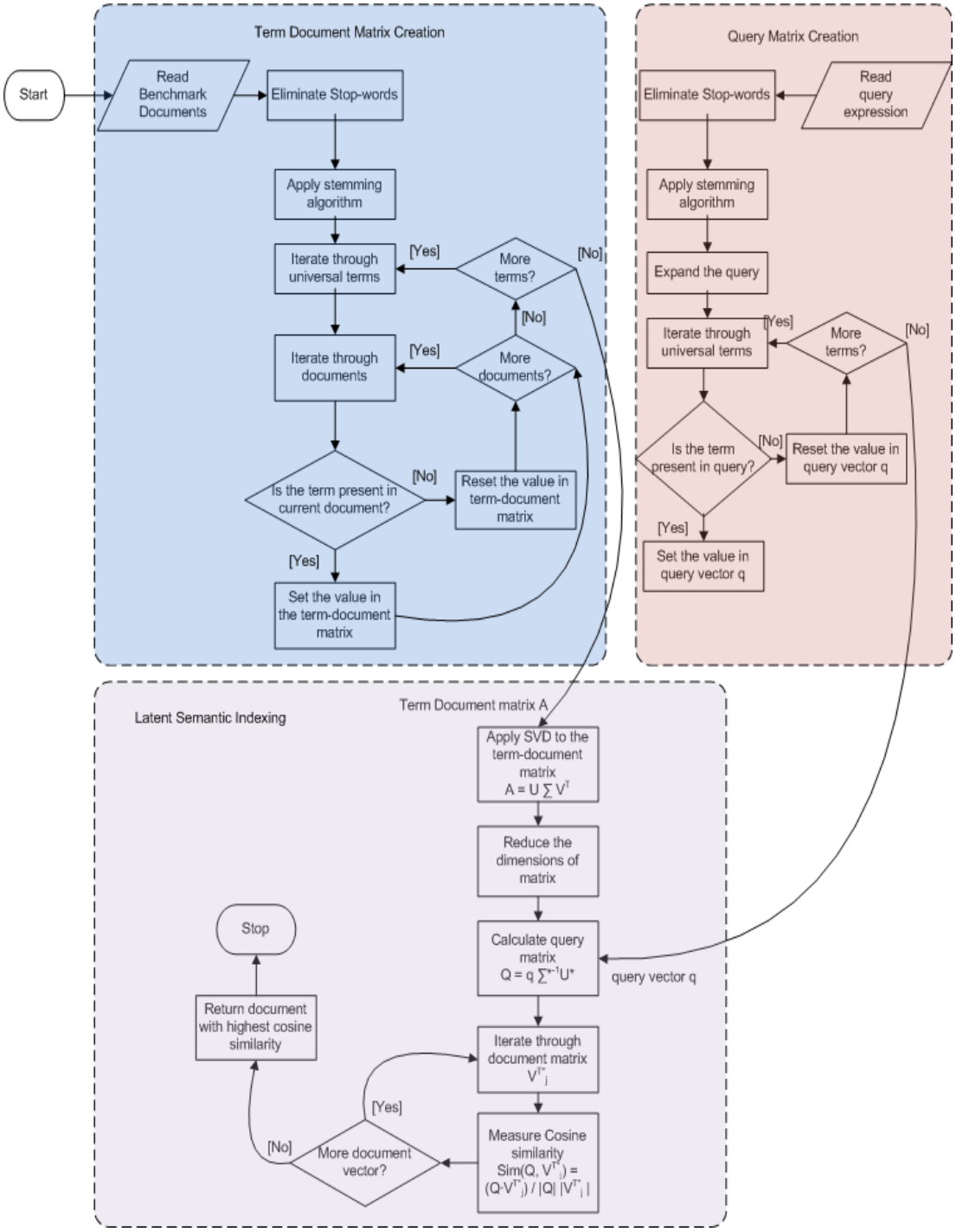

Figure 4-5 Flowchart of the Automated Nursing Knowledge Classification

4.2.1 Initial Benchmark processing using Lucene API

Apache Lucene is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform [27]. It was originally created by Dough Cutting. It has been ported to programming languages including Delphi, Perl, C#, C++, Python, Ruby and PHP[28].

The fundamental concepts in Lucene are index, document, field and term. The index is defined as a sequence of documents. The documents contain a sequence of fields. A field is a named sequence of terms. A term is a pair of strings, one indicating the name of the field while the other indicating its value.

Lucene stores the input in a data structure known as inverted index. This data structure makes efficient use of disk space as will as allows quick keyword look up. Hence instead of find out "What words are contained in this document?" it answers "Which documents contain this word?"[29]. The core of all today's web search engines is inverted index.

We use Lucene to do the initial processing of our benchmark document database. We extract the content of each document and store it under a single field named "text". Using Snowball Analyzer provided by Lucene, we will eliminate the stop words as well as apply stemming algorithm to the documents.

Below is the code snippet from our implementation.

```
 void luceneIndex()
        {
            SnowballAnalyzer a = new SnowballAnalyzer("English",
STOP_WORDS);
            indexWriter = new IndexWriter(this.pathIndex, a, true);
```

```
            bytesTotal = 0;
            countTotal = 0;
            countSkipped = 0;

            string searchPath =
Path.Combine(Environment.GetFolderPath(Environment.SpecialFolder.Person
al), "NLG");
            DirectoryInfo di = new DirectoryInfo(searchPath);

            DateTime start = DateTime.Now;

            addFolder(di);

            string summary = String.Format("Done. Indexed  files ({1}
bytes). Skipped {2} files.", countTotal, bytesTotal, countSkipped);
            summary += String.Format(" Took {0}", (DateTime.Now -
start));
            status(summary);

            indexWriter.Optimize();
            docCount = indexWriter.DocCount();
            indexWriter.Close();
        }

private void addFolder(DirectoryInfo directory)
        {
            // find all match//ing files
            foreach (FileInfo fi in directory.GetFiles("*.doc"))
            {
                // skip temporary office files
                if (fi.Name.StartsWith("~"))
                    continue;

                try
                {
                    addOfficeDocument(fi.FullName);

                    // update statistics
                    this.countTotal++;
                    this.bytesTotal += fi.Length;
                    // show added file
                    status(fi.FullName);
                }
                catch (Exception)
                {
                    // parsing and indexing wasn't successful, skipping
that file
                    this.countSkipped++;
                    status("Skipped: " + fi.FullName);
                }
            }
            // add subfolders
            foreach (DirectoryInfo di in directory.GetDirectories())
            {
```

```
                addFolder(di);
            }
        }

        private void addOfficeDocument(string path)
        {
            Document doc = new Document();
            string filename = Path.GetFileName(path);

            doc.Add(new Field("text", Parser.Parse(path),
Field.Store.YES, Field.Index.TOKENIZED));
            doc.Add(new Field("path", path, Field.Store.YES,
Field.Index.NO));
            doc.Add(new Field("title", filename, Field.Store.YES,
Field.Index.NO));
            indexWriter.AddDocument(doc);
        }
```

In this code-snippet, from line 24 to line 65 the *SnowballAnalyzer* traverses through each word document in the benchmark collection. Line 3 to line 22 eliminates the stop words from the data collected and applies the stemming algorithm to it. It also indexes the documents.But we will use the Lucene index only to get the term enumerator. The term enumerator will help us build our term document matrix.

Each document that is indexed, is made up of three fields text, path and title. Parser class extracts the content of each document into “text” field of the document. This is done by using a special filter class “IFilter” provided by microsoft. As the names suggests, “path” field holds the value for the physical path or location of the document and “title” field holds the value for the tile of the document.

4.2.2 Document Indexing

A term-document matrix is $m \times n$ matrix where m is the number or terms in the overall benchmark data and n is the number of benchmark documents. After applying

singular vector decomposition (SVD), this matrix is decomposed into three different matrices.

$$X = U \sum V^T \tag{10}$$

Here, U is a $m \times n$ right singular matrix, $\sum$ is $m \times n$ diagonal matrix and V is $n \times n$ left singular matrix. While V^T is transpose of matrix V.

The values in matrix U denote the terms involved in the benchmark database while the values in V denote the documents in the database. The diagonal elements in matrix $\sum$ represent the singular values for the term-document matrix.

Below is the code snippet from our implementation to build the term document matrix by traversing through all the terms in the Lucene Index.

```
ArrayList matrix = new ArrayList();
IndexReader reader = IndexReader.Open(this.pathIndex);
TermEnum termEnum = reader.Terms();
TermDocs termd = reader.TermDocs();
mUniversalTerms = new ArrayList();
while (termEnum.Next())
{
    if (termEnum.Term().Field() == "text")
    {
        mTermDoc = new ArrayList(4);
        for (int i = 0; i < 4; i++)
            mTermDoc.Add(0.0);
        termd.Seek(termEnum);
        while (termd.Next())
        {
            dn = termd.Doc();
            double freq = termd.Freq();
            if (termEnum.DocFreq() > freq)
                freq = Convert.ToDouble(freq) /
termEnum.DocFreq();
            mTermDoc[dn] = Convert.ToDouble(freq);
        }
        mUniversalTerms.Add(termEnum.Term());
        matrix.Add(mTermDoc.ToArray(typeof(double)));
    }
}

termdocMatrix = new GeneralMatrix(matrix.Count, 4);
```

```
            for (int x = 0; x < matrix.Count; x++)
            {
                for (int y = 0; y < 4; y++)
                    termdocMatrix.SetElement(x, y,
((double[])matrix[x])[y]);
            }
```

In line 2 of the code-snippet, the *reader* reads the terms in the Lucene Index stored at the location specified by *pathIndex*. The variable *termEnum* on line 3 is the enumerator to iterate through the terms in the index while the variable *termd on line 4* is the enumerator to iterate through the documents in the index. From line 5 to line 33 shows the creation of term-document matrix *termdocMatrix*, where each element in the matrix is set to the weight of the term. If the term appears in multiple documents, the weight of the term is frequency of the term inside document normalized by the overall frequency of the term in the collection.

We then decompose the term document matrix using Singular Vector Decomposition technique as shown in the code-snippet below:

```
SingularValueDecomposition svd = new
SingularValueDecomposition(termdocMatrix);
```

4.2.3 Truncated matrix to induce semantics

Once we decompose the term-document matrix into $U, \sum$ and V matrix, it is desirable to reduce the dimensions of the matrix by keeping first k terms. This means keeping the first k rows of $\sum$ and V^T and the first k columns of U .This will map each document in the vector space representation based on term frequency to a vector in a lower dimensional space. The claim is that, the similarity between vectors in the reduced space, for example using the cosine similarity measure may be a better retrieval indicator than

similarity measured in the original term space. This is primarily because, in the reduced space, two related documents may be represented similarly even though they do not share any keywords. This may occur, for example, if the keywords used in each of the documents co-occur frequently in other documents [30].

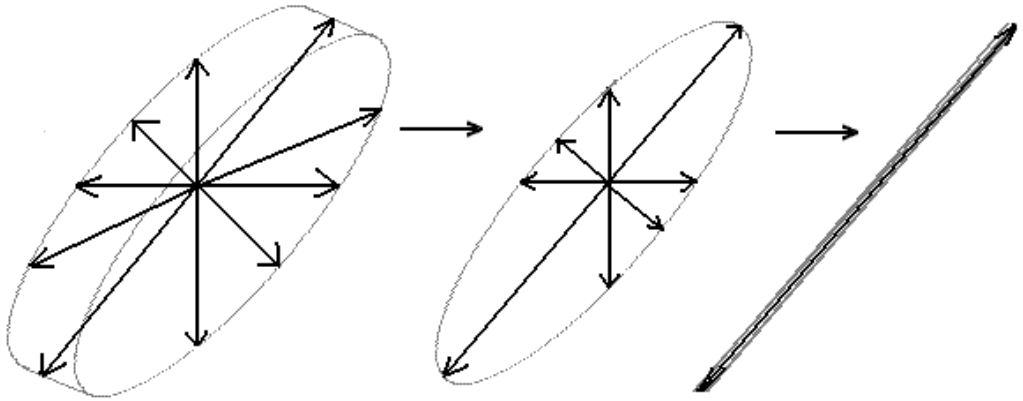

Figure 4-6 Effect of Dimensionality Reduction

Let us assume that we get three singular values from the SVD decomposition of a matrix. And we decide to keep two singular values and ignore the third singular value. This will convert the $\sum$ matrix to 2×2 matrix. Geometrically, the 3-dimensional ellipse collapses to 2-dimensional ellipse. If we keep only one singular value, the two-dimensional ellipse further collapses into a one–dimensional line. Figure 4-6 illustrates this effect of dimensionality reduction [31].

Following code-snippet shows the dimensionality reduction in our implementation:

```
GeneralMatrix U = svd.GetU();
            GeneralMatrix Sinv = svd.S.Inverse();
            documentMatrix = svd.GetV().Transpose();

            //Dimensionality reduction
            Sinv.Resize(3, 3);
            U.Resize(mUniversalTerms.Count, 3);
```

```
documentMatrix.Resize(3, 4);
```

Here, as we have four documents in the collection of our benchmark data, we decided to keep the first three rows of $\sum$ and V^T and the first three columns of U. In the code-snippet above, line 5 reduces the size on *Sinv* to 3 X 3 matrix, line 6 reduces the number of rows in matrix *U* to 3 and line 4 reduces the size of *documentMatrix* to 3 X 4.

4.2.4 Processing the query vector

Now, we prepare a query vector by inspecting the terms in the query against the terms in the overall benchmark database. As we have single sentenced queries, to achieve better results, we expand the query by including the synonyms of the query terms into the query vector.

We use *SynExpand* class provided by Lucene.net to expand our query string to include the synonyms of the terms included in the query as shown by the code-snippet below:

```
IndexSearcher syns = new IndexSearcher(directory);

            Query q1 = SynExpand.Expand(query.ToString(), syns, new
SnowballAnalyzer("English", STOP_WORDS), "text", 0.9f);
```

The final query Vector is then calculated as,

$$Q = q^T \times \sum^{*-1} \times U^* \tag{11}$$

Where,

q^T = Transpose of the query vector derived by inspecting each term in the query against universal terms.

$\sum^{*-1}$ = Moore-Penrose pseudo-inverse matrix obtained by inverting all the non-zero elements in the truncated diagonal matrix $\sum$.

U^* = Truncated matrix U.

4.2.5 Predicting the semantic closeness

We will predict the category of our query-strings by measuring how close each document vector is to the query vector. This is done by measuring the cosine angle between the two vectors. As the cosine of angle between two vectors approaches 1, it indicates that the two documents are getting closer. So the document is semantically close to the query. Similarly, if the cosine of angle approaches zero, the query and the document have no similarity between them. Figure 4-7 shows the two vectors in semantic space.

We calculate the cosine similarity between the query vector q and each document vector d_j using cosine similarity formula.

$$Sim(q, d_j) = \frac{\lfloor q \bullet d_j \rfloor}{|q| \times |d_j|} \quad \textbf{(12)}$$

Here, $(Q \bullet V^T j)$ is the dot product of the two vectors and $|Q|$, $|V^T j|$ are the Frobenius Norms of Q and $V^T{}_j$ also called Euclidean Norms which give normalized unit vectors in case of the single column matrices. The Euclidean Norm of a single column

matrix is calculated as the square root of sum of absolute square values of each element in the column.

Higher the value of Sim for the query vector and the document, more semantically similar they are. So we choose the document with the highest cosine similarity as the nursing category for the query string.

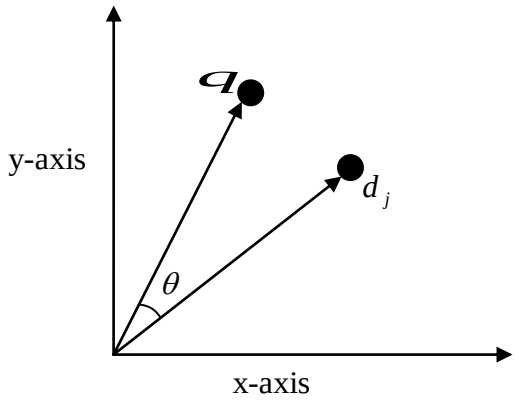

Figure 4-7 Cosine Similarity between Query Vector q and Document dj

Following is the code-snippet where we calculate the semantic closeness between the query vector and each document in the collection:

```
 double[] sims = new double[4];

 GeneralMatrix QMatrix = queryVector.Transpose() * U;
             QMatrix = QMatrix * Sinv;
             queryVector = QMatrix.GetMatrix(0, 0, 0,
QMatrix.ColumnDimension - 1);
             for (int i = 0; i < 4; i++)
             {
                 GeneralMatrix docVector = documentMatrix.GetMatrix(0,
documentMatrix.RowDimension - 1, i, i);
                 double dotprod = (queryVector *
docVector).GetElement(0, 0);
                 double queryNorm = queryVector.Norm1();
                 double docNorm = docVector.Norm1();
```

```
                double sim = (queryVector * docVector).GetElement(0, 0)
/ ((queryVector.Norm1()) * (docVector.Norm1()));
                sims[i] = sim;
            }
```

In this code-snippet, line 2 to line 5 calculates the query matrix *queryVector*, which in our case is a matrix with a single row. It uses equation 11 to calculate the query matrix. Line 6 to line 17 calculates the cosine similarity *sim* between *queryVector* and *docVector* using question 12. Array *sims* on line 16 is an array containing the cosine similarity values of each document vector in the benchmark collection against the query vector. Finally we select the document represented by the index of highest value in the *sims* array.

4.2.6 Negative feedback loop to boost the accuracy

When a part or whole of the output is fed back into the system as a part of its input, it is called a feedback. While negative feedback seeks to cancel the output that caused it and hence also known as self correcting or balancing loop [32].

We implement negative feedback mechanism in our framework by adding events to the application. We will allow the nurse to change the category of the data which she thinks is misclassified. An event will be called whenever the category of the nursing knowledge data is changes by the nurse. On such event, we will add the nursing knowledge data into the corresponding benchmark document. This will change our term-document matrix. So the next time when nurse inputs a query similar to this data, it will be correctly classified.

Following screen-shot shows the pop-up box that allows the nurse to re-map the expression to a different category.

In Figure 4-8, we are reassigning the misclassified expression to the caring category. This will automatically add the expression to the benchmark document and trigger the indexing of the documents.

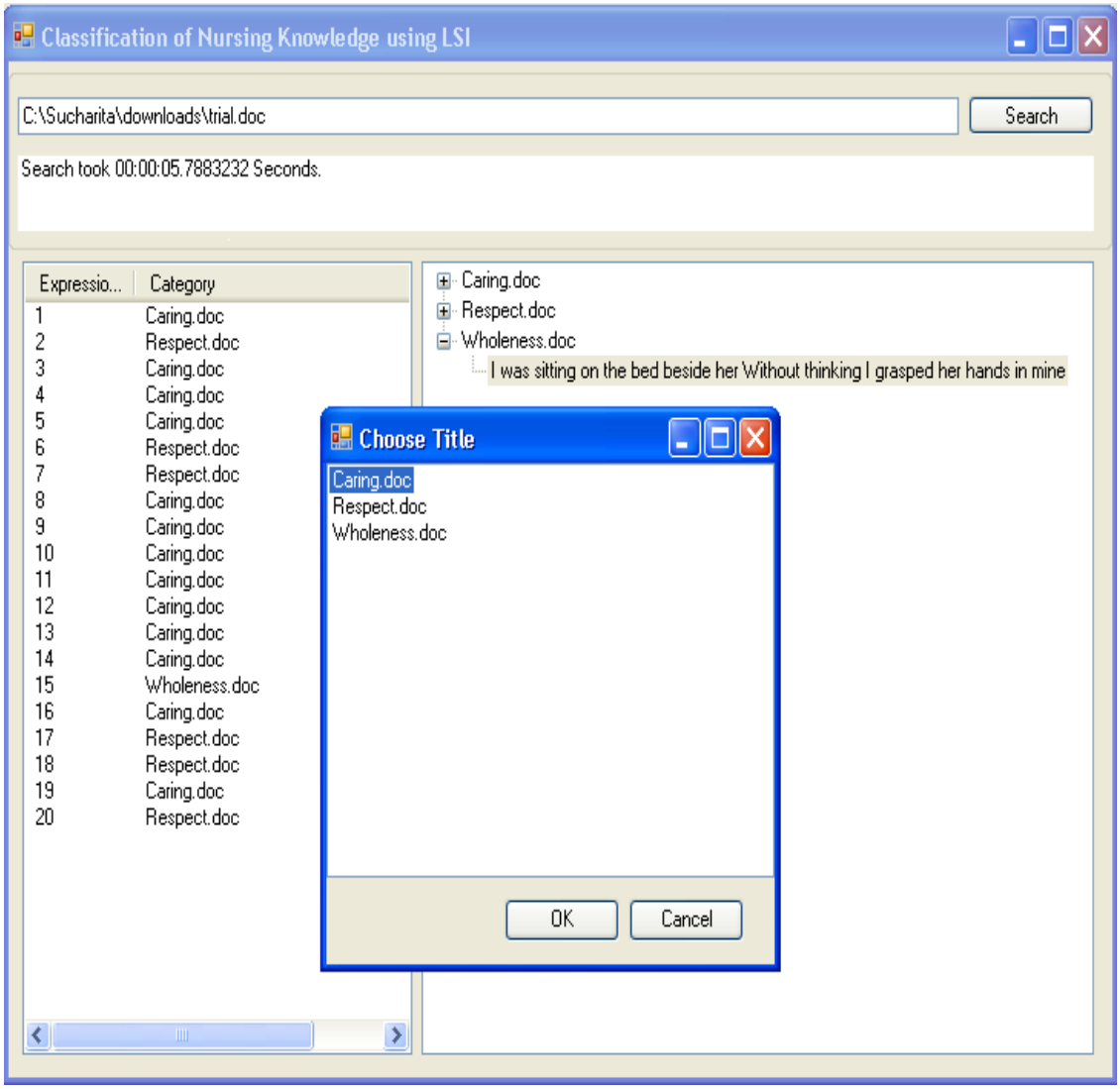

Figure 4-8 Screen Shot of the Pop-up Box for Reassigning the Expressions

4.2.7 Assumptions and limitations of using LSI

LSI understands the semantics of the nursing knowledge expressions from the terms in it after eliminating the stop words and applying stemming algorithm. As most of the nursing knowledge categories tend to talk about the care given to the patient, if we have only one or two words in the nursing knowledge expressions, there is lesser possibility of getting accurate results. In order to get accurate results, we need to have longer nursing knowledge expressions.

4.3 Simulation results and analysis

In order to test the accuracy of output of our simulation, we processed a test document containing 20 Expressions related to "Respect" category. Following are the expressions input to the simulation:

1. I respected and kept in mind the different types of cultures that I was dealing with.
2. Be courteous and respectful.
3. Allow the patient to die in a dignified and comfortable way.
4. Only two nurses believed me when I said I was in pain.
5. She called me Teacher Chen that made me feel respected.
6. I took the traditional herbal medicine, that is my belief and she respected that.
7. My nurse asked about my daily needs I am vegetarian so I need special meals.
8. The nurse drew the curtains and covered me first before changing the dressing for me She respected my privacy at all times the caring nurse would ask me where I wanted to have a shot before she did it.

9. Speak with me in a warm and caring tone of voice.
10. Respect includes esteem for family and community, as well as environment.
11. Look at me when speaking to me.
12. Believe me when I say I am in pain and when I express my personal feelings, such as feeling afraid or alone.
13. Be gentle in your actions, for example, when you help me use the bed pan or turn over.
14. Address me properly and make me feel respected.
15. Respect my privacy at all times during physical examinations, therapy, and nursing care For example, draw the curtains or cover me in an appropriate manner.
16. Respect my religious beliefs.
17. Respect my individual privacy and keep my medical information privileged.
18. Most important is that they treat me like I am a person, you know, that I am not the gray-haired lady in room.
19. Believe me when I say I am in pain and when I express my personal feelings.
20. Respect is giving them the autonomy.

Figure 4-9 shows the screenshot of the results that we got after processing the test document. Out of 20 test expressions, 19 were correctly classified to belong to the "Respect" category. So in this case the accuracy of the classification is 95%.

4.4 Result comparison

We developed an application that will use Lucene to index the benchmark documents. Lucene uses a modified vector space model for its search and not Latent Semantic Indexing. We will compare the results of this application to the application that uses LSI. Figure 4-10 shows the comparison of results for the classification of 20 expressions belonging to "Respect" category. As shown in the Figure 4-10, 19 expressions were classified correctly using LSI technique as opposed to 16 expressions using Lucene indexing technique.

The sentence that was misclassified to belong to the caring category is:

- My nurse asked about my daily needs I am vegetarian so I need special meals.

  This sentence could belong to both caring as well as respect category. So it should be mapped to both the categories. This feature of mapping an expression to multiple categories is not implemented in this application and could be added in the future.

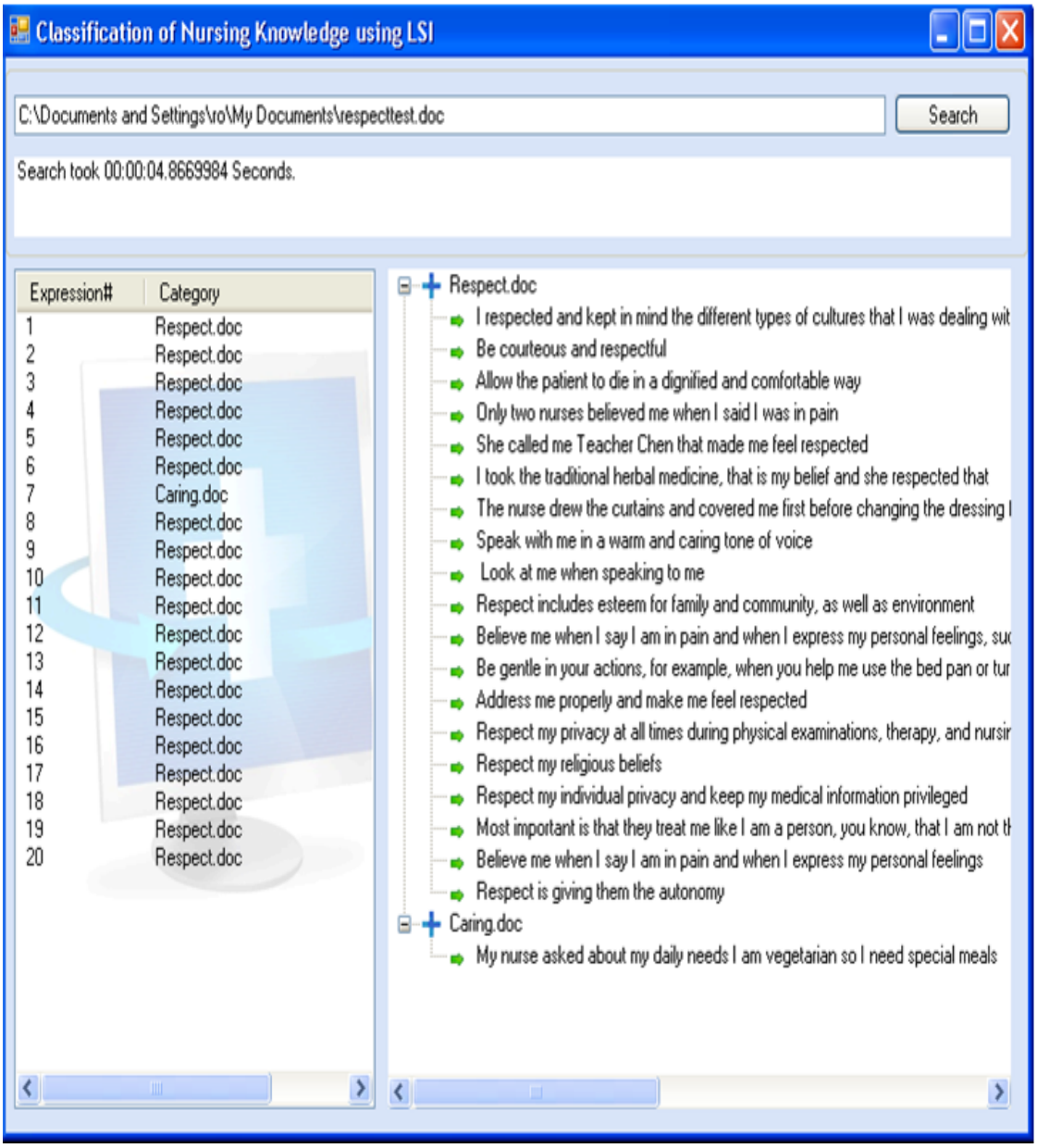

Figure 4-9 Screen Shot of the Classification of 20 Expressions belonging to the Respect Category

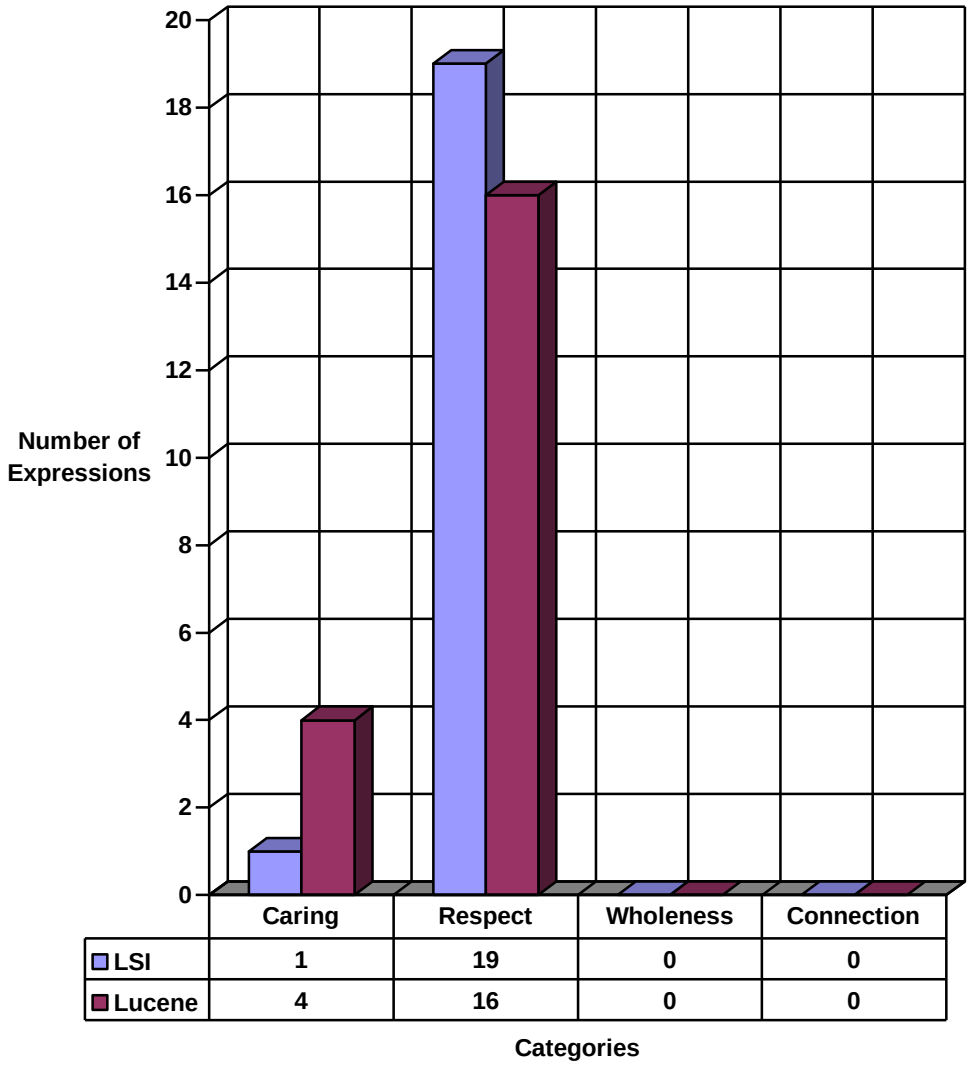

Figure 4-10 Comparison of Results for "Respect" Category

In our next test, we compared the results for 20 expressions of caring category. Figure 4-11 below shows the chart for the comparison of results for the classification of

20 expressions using LSI technique and Lucene indexing technique. For the application that used LSI technique, 16 expressions out of 20 were classified as they belong to 'caring category while three expressions were classified to belong to the 'respect' category and two expressions to the 'wholeness' category. The application using Lucene indexing technique classified 16 expressions correctly as belonging to "Caring" category while four expressions were misclassified as "Respect" category expressions.

Hence based on these observations, the accuracy of classification could be calculated.

$$AverageAccuracy = \frac{\sum_{C=1}^{T} \frac{n_C}{N_C}}{T} \times 100 \quad (13)$$

Where T = Number of categories

n_C = Number of correctly classified expressions for each category

N_C = Number of expressions in each category

Using Equation 13, the average accuracy for application using LSI and Lucene indexing techniques comes up to be 87.5% and 80% respectively.

$$\left(AverageAccuracy\right)_{LSI} = \frac{\left(\frac{19}{20} + \frac{16}{20}\right)}{2} \times 100 = 87.5\%$$

$$\left(AverageAccuracy\right)_{Lucene} = \frac{\left(\frac{16}{20} + \frac{16}{20}\right)}{2} \times 100 = 80\%$$

Figure 4-12 shows the comparison of accuracy of LSI technique and Lucene Indexing technique.

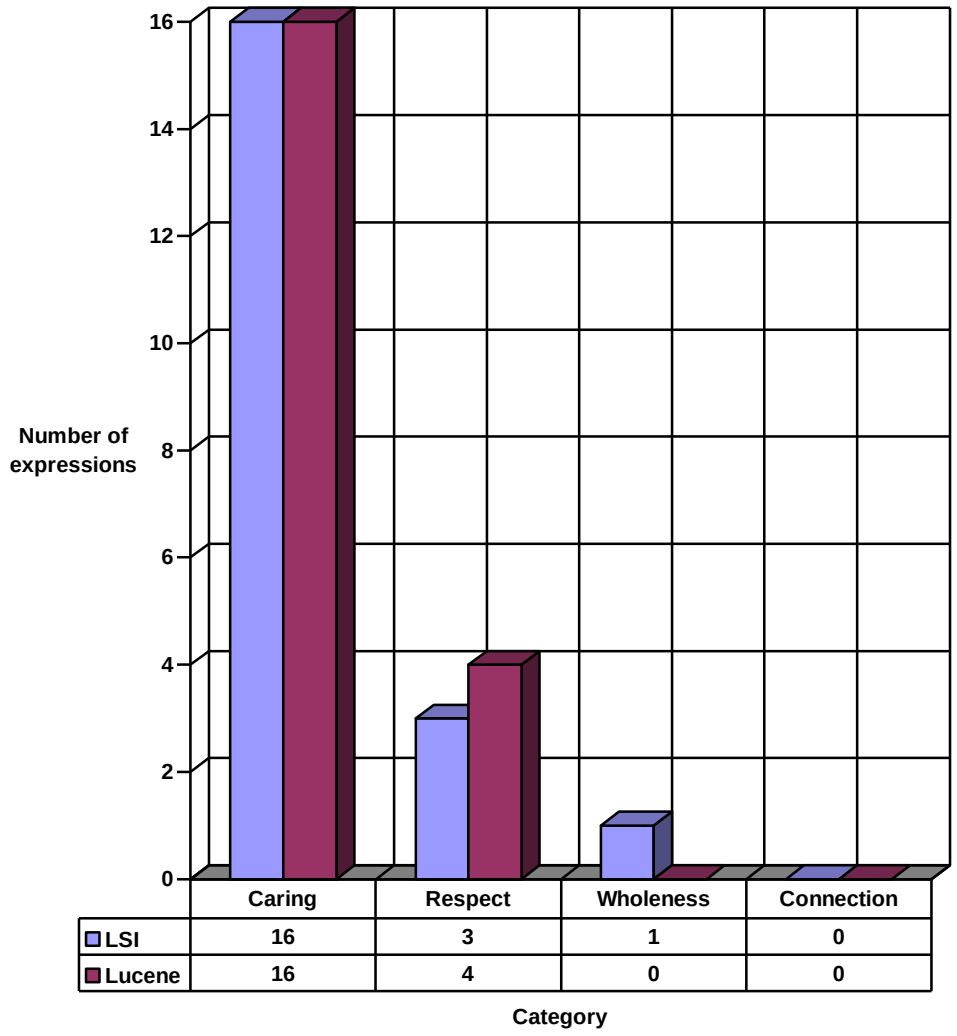

Figure 4-11 Comparison of Results for “Caring” Category

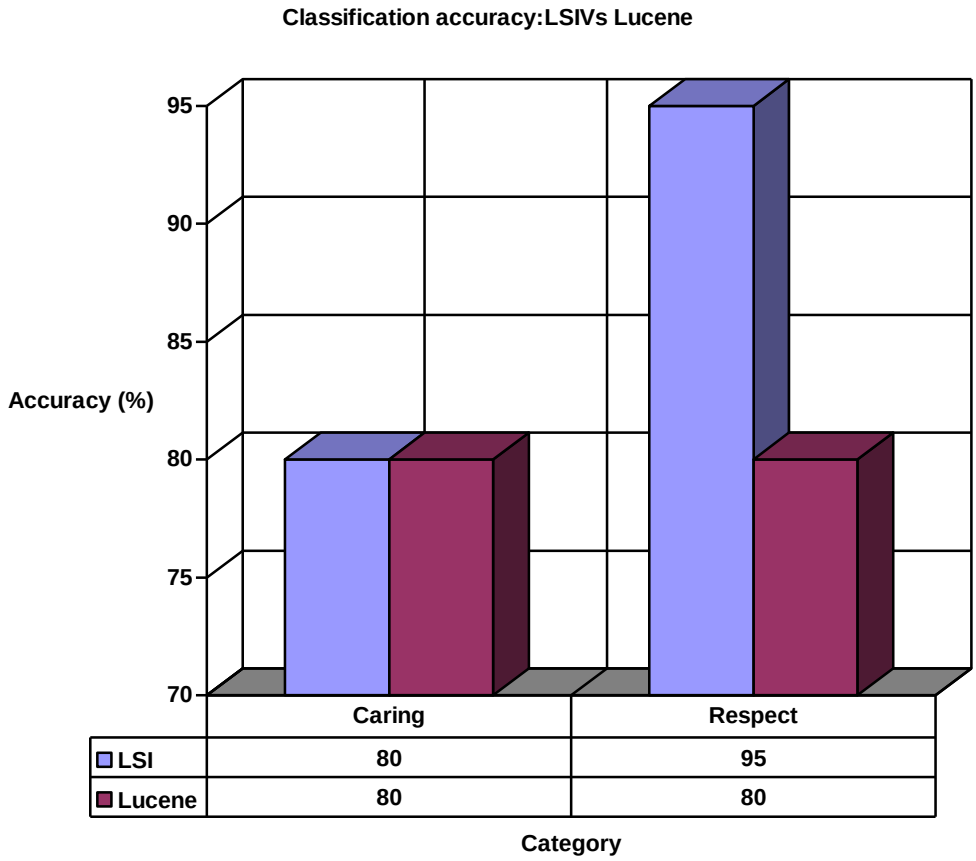

Figure 4-12 Comparison of the Classification Accuracy.

Now, the misclassified expressions, could also be mapped to the correct category be the nurse by double-clicking the expression and choosing the correct category. Figure 4-13 shows the screen-shot of this feature. Whenever an expression is re-assigned to a different category, the application copies the respective expression in the document representing the correct category in the benchmark collection. This ensures that, if a similar expression appears again in some other assessment data, the expression will not be misclassified. This feature can also be called as loopback or negative feedback feature.

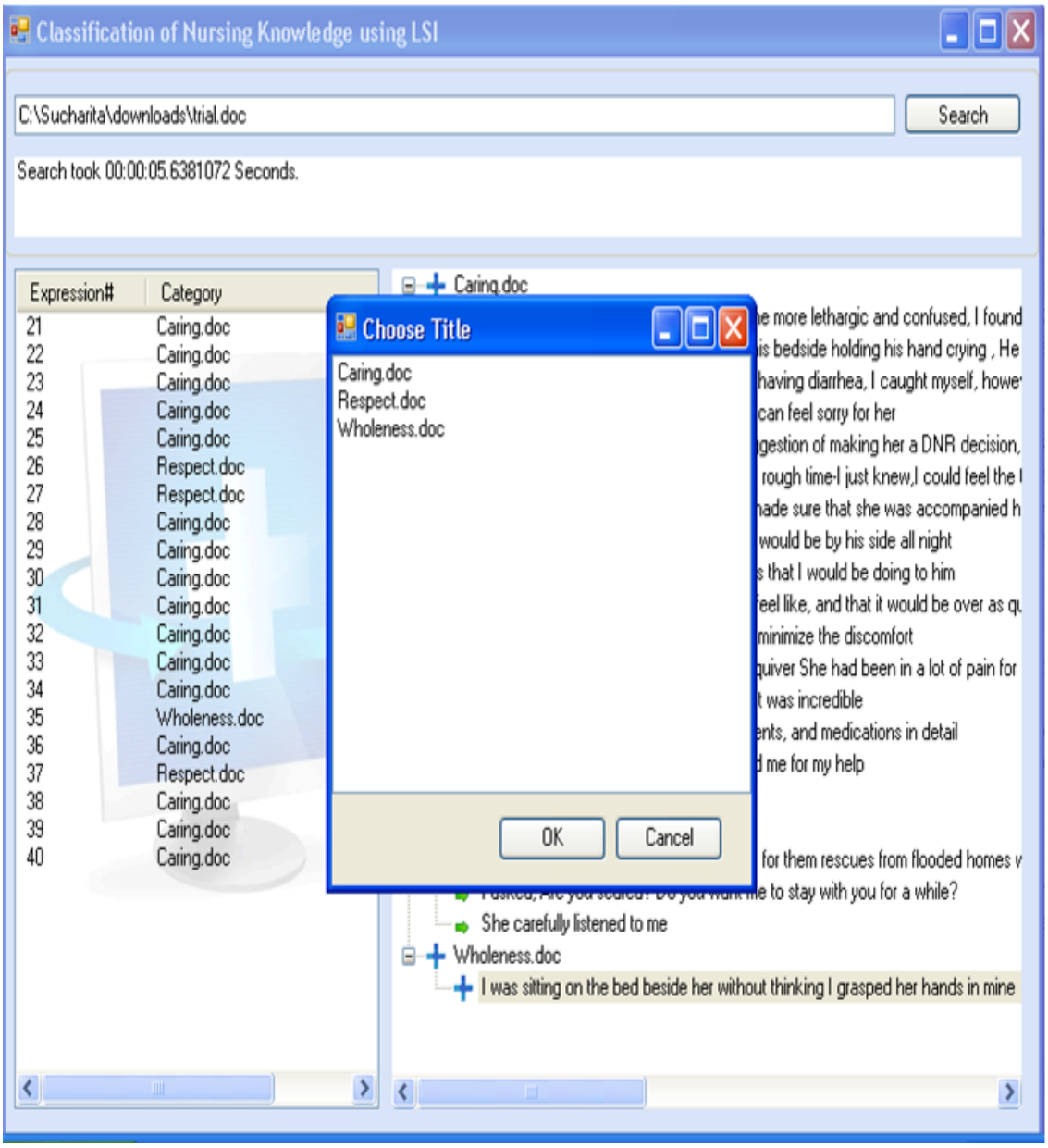

Figure 4-13 Re-assigning Misclassified Expressions to the Correct Category

4.5 Summary

In this chapter, we simulated the proposed framework in Chapter 3 using Latent Semantic Indexing technique. The nursing theory model used is "Barry/Parker

Community Nursing Practice Model". The nursing knowledge data used is a random collection of 20 expressions per category. The data for benchmark model is collected with the help of expert nurse practitioners and professors. The benchmark documents are pre-processed using classes defined by Lucene.net framework. A term-document matrix is prepared which then undergoes singular vector decomposition. We also use the dimensionality reduction technique to increase the accuracy of the results. Then each nursing knowledge expression is processed and converted into a query vector. The category for the nursing knowledge expression is predicted by computing the cosine similarity between the corresponding query vector and each document vector obtained from the SVD.A negative feedback mechanism is implemented to increase the accuracy of results. The results obtained from the simulation are compared with the results obtained from an application using Lucene indexing technique. The average accuracy of the application using LSI technique is found to be 87.5% while that of the application using Lucene indexing technique is found to be 80%.

CHAPTER 5

CONCLUSION AND SUMMARY

The world is round and the place which may seem like the end may also be the beginning.
- Ivy Baker Priest.
-

5.1 Research Summary

Complementing medical doctors and specialists, nursing care is an integral part of the holistic approach to promote health and wellbeing. The nursing knowledge captured during the interaction between nurses and patients, further with patients' families, is an invaluable part of providing a care plan and monitoring recovery progress. This thesis proposes a framework that automatically manages nursing knowledge and maps nursing practice to the caring categories according to nursing theory. The advantage of automated classification of this data is the time conservation as well as to achieve consistency in the output.

As the nursing knowledge data is in the form of natural text, we need to classify it to correct category even if the corresponding benchmark document does not have the respective keyword. Hence we use Latent Semantic Indexing (LSI) technique for classification. LSI technique helps the application to understand the semantics of the documents based on the co-occurrence of terms inside the documents.

To illustrate the validation of the framework, a case study in a nursing practice environment is presented, and the classification results are analyzed and compared with alternative approach. The result comparison shows that the LSI strategy gives 87.5% accurate results compared to the Lucene indexing technique that gives 80% accuracy. Both indexing methods maintain 100% consistency in the results.

5.2 Implications for the Nursing Community

This work will help the nursing community to design a consistent and efficient care plan for an individual patient. With a consistent and properly designed care plan, the patient will receive same quality of care even when treated by different nurses. Also it will save the amount of time and trouble for the nurse to analyze all the text assessment data and map each sentence to the correct category in the plethora of available nursing theories in the universe.

5.3 Future Work

This framework requires a clear distinction between the categories used for classification. As LSI is also known to use keyword co-occurrence, if the categories are not clearly distinguished from each other, the results are sometimes confusing. But in that case, this application could be enhanced to classify the expression to multiple categories which the LSI finds to be similar. The nurse can even eliminate the expression from the particular category if she thinks the expression is incorrectly classified.

To further automate the process of preparing the care plan for the patients, this application can be extended to allow a nurse to enter the action plan related to each classified expression to generate and save the care plan for the patient.

This application can even be used to monitor the progress of a patient throughout his treatment. As the assessment data is collected for the patient each time he visits the nurse, some special categories could be defined to track the concerns of the patient. Then the application could be modified to generate a graph showing the recovery progress of the patient based on his past and present concerns.

Sometimes in addition to the medicines prescribed, the care provided by the nurses help in the speedy recovery of the patient. The analyses of the care plan specified by the nurse on particular visit of the patient and the related graph of patient's recovery progress will further help nurses to understand the factors that helped the patient in his speedy recovery.

BIBLIOGRAPHY

[1] B. A. Swan, N. M. Lang, and A. M. McGinley, “Access to quality care: Links between evidence, nursing language, and informatics” , *NURSING ECONOMICS*, vol. 22, no. 6, pp. 2, November-December 2004.

[2] S. C. Delaune, and P. K. Ladner, *Fundamentals of Nursing: Standards and Practices*, pp. 132, Albany, NY: Thomson Delmar Learning, 2002.

[3] H. K. SIPES, "Holistic Health Care in Nursing", "http://clearinghouse.missouriwestern.edu/manuscripts/22.asp?logon=&code=", 1998].

[4] "Nurse Practitioner", 24 Aug, 2008; http://www.nursesource.org/practioner.html.

[5] B. A. Swan, N. M. Lang , and A. M. McGinley, “Access to quality care: Links between evidence, nursing language, and informatics”, *NURSING ECONOMICS*, vol. 22, no. 6, pp. 5, November-December 2004.

[6] M. E. Parker, *Nursing Theories and Nursing Practice*: Philadelphia F.A. Davis Company, 2005.

[7] A. I. Meleis, *Theoretical nursing; Development and progress*, Philadelphia: Lippincott Williams & Wilkins, 1997.

[8] M. Silva, *Philosophy, theory, and research in nursing: A linguistic journey to nursing practice*, 1997.

[9] M. M. Leininger, “Leininger's Theory of Nursing: Cultural Care Diversity and Universality”, *Nursing Science Quarterly,* vol. 1, pp. 152-160, Nov 1988.

[10] "Theory of Group Power within Organizations", Dec 2008; http://theoryofnursinggrouppower.googlepages.com/theoryofnursinggrouppower2

[11] "The Comfort Line", Dec 2008; http://www.thecomfortline.com/.

[12] K. Ferreira, "Katharine Kolcaba's Theory Of Comfort", http://www.oppapers.com/essays/Katharine-Kolcabas-Theory-Comfort/64802, [Dec 2008, 2004].

[13] M. C. Dorothy, “Perspectives of pure science”, *Nursing Research,* vol. 17, no. 6, pp. 497 - 501, 1968.

[14] J. Dickoff, P. James, and E. Wiedenbach, “Theory in a practice discipline: Part I. Practice oriented theory”, *Nursing Research,* vol. 17, no. 5, pp. 415 - 434, 1968.

[15] P. L. Chinn, and M. K. Kramer, *Integrated Knowledge Development in Nursing*, St Louis: Mosby, 2004.

[16] S. Gray J. & Forsstorm, "Generating theory for practice: the reflective technique", *Towards a discipline of nursing*, R. Gray J. & Pratt, ed., Melbourne: Churchill Liningstone, 1991.

[17] T. Muhr. "ATLAS.ti: Quality Software Developed in Germany since 1993", Sept 2008; http://www.atlasti.com/aboutUs.html.

[18] "Ontology", 0ct 2008; http://en.wikipedia.org/wiki/Ontology_(computer_science).

[19] S. B. Palmer, “The Semantic Web: An Introduction”, 2001.

[20] H. Chengiun Liu; Wechsler, “A unified Bayesian framework for face recognition”, In *IEEE Signal Processing Society 1998 International Conference on Image Processing*, 1998, pp. 4-7.

[21] B. C. Q. L. Z. Tang;, “A Clustering Based Bayesian Network Classifier”, In *IEEE International Conferenceon Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, Aug 2007, pp. 444-448.

[22] P. Foltz, "Using Latent Semantic Indexing for Information Filtering." pp. 40-47.

[23] V. Mitra, C.-J. Wang, and S. Banerjee, "A Neuro-SVM Model for Text Classification using Latent Semantic Indexing." pp. 564-569.

[24] J.-T. Sun, Z. Chen, H.-J. Zeng *et al.*, "Supervised Latent Semantic Indexing For Document Categorization."

[25] T. Scott, “The Canonical Activities of Reverse Engineering”, In *Annuals of Software Engineering*, 2000, pp. 1-4.

[26] "Stemming Algorithm", Oct 2008; http://en.wikipedia.org/wiki/Stemming.

[27] "Apache Lucene", Oct 2008; http://lucene.apache.org/java/docs/.

[28] "Lucene - Wikipedia", Oct 2008; http://en.wikipedia.org/wiki/Lucene.

[29] H. Erik, and G. Otis, *Lucene in Action (In Action series)*: Manning Publications Co., 2004.

[30] T. B. Brian, W. C. Garrison, and K. B. Richard, “Latent semantic indexing is an optimal special case of multidimensional scaling”, In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, Copenhagen, Denmark, 1992.

[31] "SVD and LSI Tutorial 1: Understanding SVD and LSI", Oct 2008; http://www.miislita.com/information-retrieval-tutorial/svd-lsi-tutorial-1-understanding.html.

[32] P. M.Senge, *The Fifth Descipline:The Art and Practice of the Learning Organization*, pp. 424, New York: Doubleday, 1990.