PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Sandeep Mudabail Raghuram

Entitled Bridging Text Mining and Bayesian Networks

For the degree of Master of Science

Is approved by the final examining committee:

Dr. Yuni Xia
_____
Chair

Dr. Mathew Palakal

Dr. Xukai Zou

To the best of my knowledge and as understood by the student in the *Research Integrity and Copyright Disclaimer (Graduate School Form 20)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): Dr. Yuni Xia

_____

Approved by: Dr. Shiaofen Fang   4/1/2010
               Head of the Graduate Program   Date

PURDUE UNIVERSITY
GRADUATE SCHOOL

Research Integrity and Copyright Disclaimer

Title of Thesis/Dissertation:

Bridging Text Mining and Bayesian Networks

For the degree of ___Master of Science_____

I certify that in the preparation of this thesis, I have observed the provisions of *Purdue University Teaching, Research, and Outreach Policy on Research Misconduct (VIII.3.1)*, October 1, 2008.*

Further, I certify that this work is free of plagiarism and all materials appearing in this thesis/dissertation have been properly quoted and attributed.

I certify that all copyrighted material incorporated into this thesis/dissertation is in compliance with the United States' copyright law and that I have received written permission from the copyright owners for my use of their work, which is beyond the scope of the law. I agree to indemnify and save harmless Purdue University from any and all claims that may be asserted or that may arise from any copyright violation.

Sandeep Mudabail Raghuram
_____
Printed Name and Signature of Candidate

04/28/2010
_____
Date (month/day/year)

*Located at http://www.purdue.edu/policies/pages/teach_res_outreach/viii_3_1.html

BRIDGING TEXT MINING AND BAYESIAN NETWORKS

A Thesis

Submitted to the Faculty

of

Purdue University

by

Sandeep Mudabail Raghuram

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

August 2010

Purdue University

Indianapolis, Indiana

To my mom, dad and sister.

ACKNOWLEDGMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GLOSSARY

BN - Bayesian Network

CPT - Conditional Probability Table

D-map - Dependency map

I-map - Independency map

ISI - Institute for Scientific Measure

IF - Impact Factor

WCNB - Weight-normalized Complement Naïve Bayes

NP - Noun Phrases

ABSTRACT

Raghuram, Sandeep Mudabail. M.S., Purdue University, August, 2010. Bridging Text Mining and Bayesian Networks. Major Professor: Yuni Xia.

After the initial network is constructed using expert's knowledge of the domain, Bayesian networks need to be updated as and when new data is observed. Literature mining is a very important source of this new data. In this work, we explore what kind of data needs to be extracted with the view to update Bayesian Networks, existing technologies which can be useful in achieving some of the goals and what research is required to accomplish the remaining requirements. This thesis specifically deals with utilizing causal associations and experimental results which can be obtained from literature mining. However, these associations and numerical results cannot be directly integrated with the Bayesian network. The source of the literature and the perceived quality of research needs to be factored into the process of integration, just like a human, reading the literature, would. This thesis presents a general methodology for updating a Bayesian Network with the mined data. This methodology consists of solutions to some of the issues surrounding the task of integrating the causal

associations with the Bayesian Network and demonstrates the idea with a semi-automated software system.

CHAPTER 1. INTRODUCTION

1.1. <u>Objectives</u>

The overall aim of this research was to find a methodology to update Bayesian networks as and when new data is observed. Literature mining is a very important source of this new data after the initial network is constructed using the expert's knowledge. But the task of reading through hundreds of journal articles and publications to support existing associations and probabilities can become very tedious. Automated systems are not yet available because the state of the art is not developed enough to deduce associations and probabilities from a discourse level analysis of the literature article.

However, existing research demonstrates ways to extract intra-sentential causal associations. This research has explored ways to use these causal associations and the issues related to integrating it with existing Bayesian Network. This research also tries to define what kind of data in the literature can be interesting from the perspective of updating a Bayesian Network.

A human reading a literature piece, would usually associate some kind of trust or confidence in the article. This confidence could stem from the reputation of the publication house, the author of the article etc. This degree of confidence plays an important role in the reader's acceptance of the data and ultimately the data represented in the Bayesian Network. For example, if articles from two different authors and publications propose the same causal association but with different probabilities, the reader needs to make a decision as to which article to 'trust'

and what probability to use in the Bayesian Network. In this research, we make an attempt to address this issue.

The specific objectives were to:

1.   Identify the type of information in text which can be potentially useful for constructing or updating a Bayesian network.

2.   Develop a methodology to utilize the mined information.

3.   Create a semi-automated tool to demonstrate the methodology and provide the user with useful information to update Bayesian networks.

1.2. <u>Organization</u>

This thesis has 8 chapters and is organized as follows: Chapter 2 provides information about the background of this research and related work done in this field. Chapter 3 presents the outline of the methodology proposed. Chapter 4, 5 and 6 discuss each phase of the proposed in detail. Chapter 7 presents the experimental system developed and the results. Chapter 8 concludes the work. The appendix provides information about the software system developed to demonstrate the ideas proposed.

CHAPTER 2. PRELIMINARIES

2.1. <u>Background</u>

2.1.1. Bayesian Network

A Bayesian network (BN) is a directed acyclic graph whose arcs denote a direct causal influence between parent nodes (causes) and children nodes (effects) [11]. A BN is often used in conjunction with statistical techniques as a powerful data analysis tool. While it can handle incomplete data and uncertainty in a domain, it can also combine prior knowledge with new data (evidence) [4]. A BN makes predictions using the conditional probability distribution tables (CPT). Each node in a BN has a CPT which describes the conditional probability of that node, given the values of its parents [13]. Using the CPT for each node, the joint probability distribution of the entire network can be derived by multiplying the conditional probability of each node. Probabilistic inference in a Bayesian network is achieved through evidence propagation. Evidence propagation is the process of efficiently computing the marginal probabilities of variables of interest, conditional on arbitrary configurations of other variables, which constitute the observed evidence [14].

2.1.2. Constructing a Bayesian Network

Causality denotes a necessary relationship between one event ("cause") and another event ("effect") which is the direct consequence of the first [7]. It implies a dependency between a cause and an effect where the probability of the "effect" occurring becomes very high, if the "cause" occurs first in a chronological order

[1]. A causal model is an abstract model that uses cause and effect logic to describe the behavior of a system [8]. This model can then be used to build a BN. This approach of building a BN from causal modeling is essential in understanding the problem domain and predicting the consequences of an intervention [4].

There are two approaches to construct a BN: knowledge-driven and data-driven. The knowledge-driven approach involves using an expert's domain knowledge to derive the causal associations. The data driven approach uses the causal modeling technique described before, to derive the mappings from data which can then be validated by the expert [9].

2.2. Analysis of the Problem

This thesis studies the problem of updating a Bayesian Network. As discussed earlier, a BN is built initially by an expert drawing upon his/her domain knowledge. Some of this knowledge can be axiomatic, i.e accepted facts of the domain that are not expected to change over time, while the rest is mostly the belief at that point in time. This belief needs to be reinforced over time or subjected to modifications. The modifications could result in re-configuration of the causal mappings, like addition/deletion, or it could be a change in the probability. A popular implementation of BN, Netica, provides a function to 'fade' the probability associated with causal mappings in the network. This results in a reduction in the belief associated with the mapping, if it is not reinforced from time to time citing new evidences.

Case files can be a very good source of evidence. The case files might contain interventions suggested by the Bayesian Network and could provide vital information about the success or failure of those interventions. Literature is another important source of new evidences. It could be new research publication,

survey of articles in the domain or an analysis of cases and interventions for the domain. However, procuring these new evidences from literature is a tedious task. In many cases, it involves manual readings of articles and journals and manual update tasks to keep the model updated. Automated techniques exist to mine information from literature. But they are limited in scope due to the fact that text mining technology has not progressed enough to 'deduce' the meaning implied over multiple sentences, paragraphs or across the entire article. Intra-sentential mining is, however, a developed technology, with substantial theoretical framework to implement a system.

Building on this, what is required is an approach to associate a degree of confidence in the mined information. This can be viewed as an emulation of human behavior when faced with a new piece of information. An expert, reviewing literature in the domain, would implicitly associate some sort of confidence in the information, based on prior experience with the source of the article or the nature of work, as can be perceived from it.

Finally, the task of integrating the new information with the Bayesian Network needs to be addressed. Research in this area has identified several modeling issues [15].

2.3. Related Work

Mining causal associations from text using lexico-syntactic analysis has been studied in previous work [2, 3]. In [2], a method was developed for automatic detection of causation patterns and semi-automatic validation of ambiguous lexico-syntactic patterns that refer to causal relationships. This procedure requires a set each of causation-verbs and nouns frequently used in a given domain. Using these sets, all patterns of type <NP1 cause_verb NP2>, where NP1, NP2 are noun phrases, can be extracted. The authors of the above said work have used the causal verbs that they found to be the most frequent and

less ambiguous such as *lead (to), derive (from), result (from)*, etc. Some of the causal patterns identified by their system are: "Anemia are *caused by* excessive hemolysis", "Hemolysis is a *result of* intrinsic red cell defects", and "Splenic sequestration *produces* anemia". In [24], a system was also developed for acquiring causal knowledge from text.

This thesis builds on the previous work and designs a general framework for building a Bayesian network based on text mining. It tries to bring together numerous existing ideas and some new ideas in an attempt at bridging the two technologies. This complicated process is broken down into several stages and the major issues that need to be solved at each stage are discussed with possible solutions.

CHAPTER 3. ANALYSIS OF THE PROBLEM

3.1. <u>Outline of the Approach</u>

Existing text mining techniques can deliver causal associations from within a sentence or from sentences in close proximity of each other. As discussed in the previous chapter, these causal associations can be used to model the system and can be easily transformed into a Bayesian Network. Thereby, phrases containing causal associations form the most interesting data in the literature from the perspective of this work. Building on from here, techniques are required to estimate the probabilities for these associations. Further on, formal techniques are required to define and quantify the degree of confidence of the mined data. Once all of this data is available, integration issues need to be dealt with before the data can find its way into the Bayesian Network. For example: A causal map depicts causality between variables, i.e.it implies dependence between those variables. Hence it is a D-map. BNs, on the other hand, are I-maps: given a sequence of variables, an absence of arrow from a variable to its successors in the sequence implies conditional independence between the variables. Other modeling issues include:

- Eliminating circular relations
- Reasoning underlying the link between concepts
- Distinction between direct and indirect relations

This thesis, proposes a general methodology to bridge text mining and Bayesian network.

3.2. <u>The Proposed Methodology</u>

The problem of mining and integrating data into Bayesian Network can be solved in a systematic way as follows:

1. The causal associations need to be identified and extracted out of literature.

2. Any numerical data supporting these mappings needs to be extracted: The numerical data, usually percentages, decimals and numbers representing quantity, could indicate the probability of occurrence of the causal events, conditional or prior probabilities.

3. The source of the article, such as the journal, publication house, website etc, is identified and the degree of its influence in the domain under consideration is identified.

4. The perceived quality of the research is then quantified by categorizing the nature of the work and the quality of the experiments conducted to justify the claims made.

5. The confidence of the mined data is then quantified based on the measurements from steps 3 and 4.

6. Using the data from steps 2 and 5, the derived probability for the causal association from step 1 is computed.

7. The destination of the causal associations needs to be identified.

8. The causal associations need to be checked for consistency and validity with the existing network. This is a semi-automated technique and provides useful information to the human expert to perform the key decisions in the final leg of integrating the mined data.

Each of these steps is discussed in detail in the coming chapters.

CHAPTER 4. MINING CAUSAL ASSOCIATIONS

4.1. Extracting Causal Associations

Since the relation between parent and child nodes in a Bayesian Network is a cause-effect relationship, the most relevant pattern that needs to be mined is cause-effect pattern or causal patterns. Causal patterns can occur in the following ways:

- Cues such as connectives: "the manager fired John **because** he was lazy"
- Verbs: "smoking **causes** cancer"; or
- NPs: "Viruses are **the cause of** neurological diseases".

As discussed in [24], the first step in mining these patterns is identifying section of the text containing them. The next step is to analyze them by considering the presence of various connectives like conjunction, disjunction and negation. Conjunctions are better viewed as unit causes/effects, whereas disjunctions and conjunctions should be decomposed [24]. Going by this logic, a conjunction like "Corruption and insecurity" should be treated as a single event, whereas "Bacteria, germs or virus" should be decomposed into three separate atomic causal patterns, each of which contributes to the estimation of a separate conditional probability in the specification of the Bayesian network.

4.2. Extracting Probability

Once the associations are extracted, the expert is subjected to a structured interview to resolve the biases in the causal maps or given an adjacency matrix representation of the associations to specify the relations. Three direct response-

encoding methods to derive probabilities for the causal associations are described in [16]. In these methods, a subject responds to a set of questions either directly by providing numbers or indirectly by choosing between simple alternatives or bets. These are manual encoding techniques which require the knowledge and judgment of a human subject to elicit probabilities.

It might, however, be possible to develop an automated technique to augment these manual encoding procedures. The aim of this technique is to search for and utilize numerical data accompanying the sentences containing the causal associations and present it to the expert.

Percentages are a common way of summarizing a statistical result. Sentences containing a causal association might also contain percentages from surveys and experiments to emphasize the relation. Hence, it is useful to examine sentences marked as containing causal associations for numerical details, which can yield statistical data for the BN. It can be observed that a percentage usually occur in close proximity of the noun phrases, which are part of a causal relationship. Simple sentential structures include:

<*numerical_string_pre* NP1 causal_verb NP2>
<NP1 causal_verb NP2 *numerical_string_post*>
Where:
numerical_string_pre, *numerical_string_post* can be "*xx%*","*xx% of*","*xx% of the times*" etc

For example: "*20%* falls lead to death", "*5% of* people who fall require hospitalization", "*25% of the time* fall can result in fracture", "Falls can result in fracture *25% of the times*" etc. This percentage value can then be directly converted to the probability value for that assertion.

The strength of a causal association in text can also be estimated by looking for superlatives and other phrases which qualify the verb. For example: "There is a *strong possibility* that falls result in fracture". A list of such phrases can be mapped to pre-defined probability values.

While these patterns yield the probabilities or causal strength of the relations, other intra-sentential patterns might yield prior-probabilities for nodes in a BN. For example: "*In the age 65-and-over population as a whole, approximately 35% to 40% of community-dwelling, generally healthy older persons fall annually.*" In the domain of Geriatry, the population of interest are always persons 65 years of age or older. Under that assumption, the above sentence would yield the prior probability for a node 'fall' in a BN for 'fall risk', a prior probability of 0.375 (average). Now if the literature contained another sentence like "*55% of the people above the age of 80 were at the risk of falling*", then the two sentences put together would yield conditional probabilities for continuous valued nodes named 'age' for the ranges 65<= age < 80 and 80 <= age. This would require the knowledge of population distribution for the two age groups which would then be considered their prior probability.

However, this topic can be a subject for future research and is not addressed in this work.

CHAPTER 5. DEFINING THE CONFIDENCE MEASURE

5.1. <u>Parameters Considered</u>

One of the main focus areas of this research has been a method to determine how much confidence can be associated with the causal associations mined from text. The confidence measure is a score we associate with every causal mapping in the BN based on the confidence we have in asserting that relationship. It is an attempt at quantifying the confidence placed in the causal relationship uncovered by automated methods. This confidence stems from two sources:

- The literature source
- The nature and perceived quality of the work which puts forth the causal relation (or evidence from the perspective of the Bayesian Network)

We attempt to quantify these two sources in order to derive a formal 'confidence' measure. Hence, the two sources will be referred to as the journal's influence measure and the evidence level of the evidence.

5.1.1. Quantifying the Influence Measure

Various measures have been suggested for measuring a journal's influence. The most commonly used ones are Institute for Scientific Information (ISI) Impact Factor [18] and Eigenfactor.

The impact factor, often abbreviated IF, is a measure of the citations to science and social science journals. It is frequently used as a proxy for the importance of a journal to its field [12]. The impact factor of a journal is calculated based on a

two-year period. It can be viewed as the average number of citations in a year given to those papers in a journal that were published during the two preceding years.

For example, the 2003 impact factor of a journal would be calculated as follows:

$$IF = A / B \qquad \text{Eq. 5.1}$$

Where,

A = the number of times articles published in 2001-2 were cited in indexed journals during 2003

B = the number of "citable items" published in 2001-2

PageRank is a link analysis algorithm used by the Google Internet search engine that assigns a numerical weighting to each element of a hyperlinked set of documents [19]. The algorithm may be applied to any collection of entities with reciprocal quotations and references, such as articles published by a journal. A version of PageRank has been proposed as a replacement for the ISI impact factor, called Eigenfactor [17]. In this measure, journals are rated according to the number of incoming citations, with citations from highly-ranked journals weighted to make a larger contribution to the Eigenfactor than those from poorly-ranked journals [20].

A third way would be for a domain expert to manually assign influence measure for the journals in the domain. But such a process is not only time consuming, but could also be tedious for domains which have a large number of publishing journals. Moreover, the task of keeping this measure updated also becomes very tedious.

The final choice of the influence measure depends on the expert.

5.1.2. Quantifying the Evidence Level

Evidence level refers to a categorization or ranking of the evidence. This is a domain specific qualification of the evidence. Medicine is one domain where professionals and experts actively review literature to stay updated with current trends in treatment and best practices. Also, it is a domain where vast amounts of research and scholarly articles are published regularly in journals and websites. As a result, significant research has also been done into how to assess these large quantities of information forth coming every day. Evidence Based Medicine or EBM as it is called, is a result of this effort at categorizing evidences into qualitative levels. Evidence-based medicine categorizes different types of clinical evidence and ranks them according to the strength of their freedom from the various biases that beset medical research [10]. It also lists some commonly used evidence categories.

In general, a scheme for categorizing evidences needs to be developed for the domain under consideration. This categorization technique can then be applied in conjunction with text mining to quantify the strength of the evidence discovered.

5.1.3. Estimating the Evidence Level

Estimating the evidence level requires keyword search and/or semantic analysis of the document title, abstract, conclusion and the segment of the text containing the sentence with the causal associations. For example, in Geriatric evidence based practice, [23] lists the levels of quantitative evidence from 1 to 6, in descending order of importance. Documents containing a level-2 evidence usually have the string "Randomized Control Trial" mentioned either in their title, abstract or keywords section. However, a more detailed discussion of this topic is necessary and will not be addressed as part of this thesis.

The evidence level is then mapped to a value between [0, 1], which can be used in a formula to compute the confidence measure. In Geriatrics, level 1 corresponds to the most trusted and will hence get the highest value assigned, in this case a value of 1.

<div align="center">5.2. <u>Format for Extracting Data</u></div>

Based on the theory presented till here, the format for representing data mined from literature is shown in Table 5.1. We assume that by using the existing techniques, causal associations are extracted and available in the format.

<div align="center">Table 5.1 Format of Extracted Data</div>

Noun Phrase 1	Causal verb	Noun Phrase 2	Probability	Evidence Level

Noun phrase1, causal verb, Noun phrase2 represent the triplet mined from text using techniques mentioned above. Causal verb is not a mandatory field but may be useful in identifying the directionality of the relationship i.e. it may help in identifying if Noun Phrase 1 is the cause or the effect. It is useful to differentiate triplets like: "Slippery road is caused by snow.", "Slippery roads cause accidents." In the absence of this field, it is assumed that NP1 is the cause and NP2 is the effect.

Probability is the prior probability for the causal mapping, which can be extracted from text using additional semantic analysis or assigned a default value.

Consider the following sentence:
"*For persons age 65 and older, 25% of falls result in fracture*"

It can be decomposed as shown in Table 5.2.

Table 5.2 Example of Extracted Data

falls	result in	fracture	0.25	Level 2

5.3. Derive the Confidence Measure

The chosen influence measure for the domain is normalized to a value [0, 1] for every journal. The confidence measure is then computed as a weighted average of these two parameters:

$$\text{confidence\_measure} = \frac{\left(\left(w_i * \text{influence\_measure}\right) + \left(w_e * \text{evidence\_level}\right)\right)}{\left(w_i + w_e\right)} \quad \text{Eq. 5.2}$$

Here W_i and W_e are the weights assigned to influence measure and evidence level respectively. Their values will be determined at the expert's discretion and could vary from domain to domain.
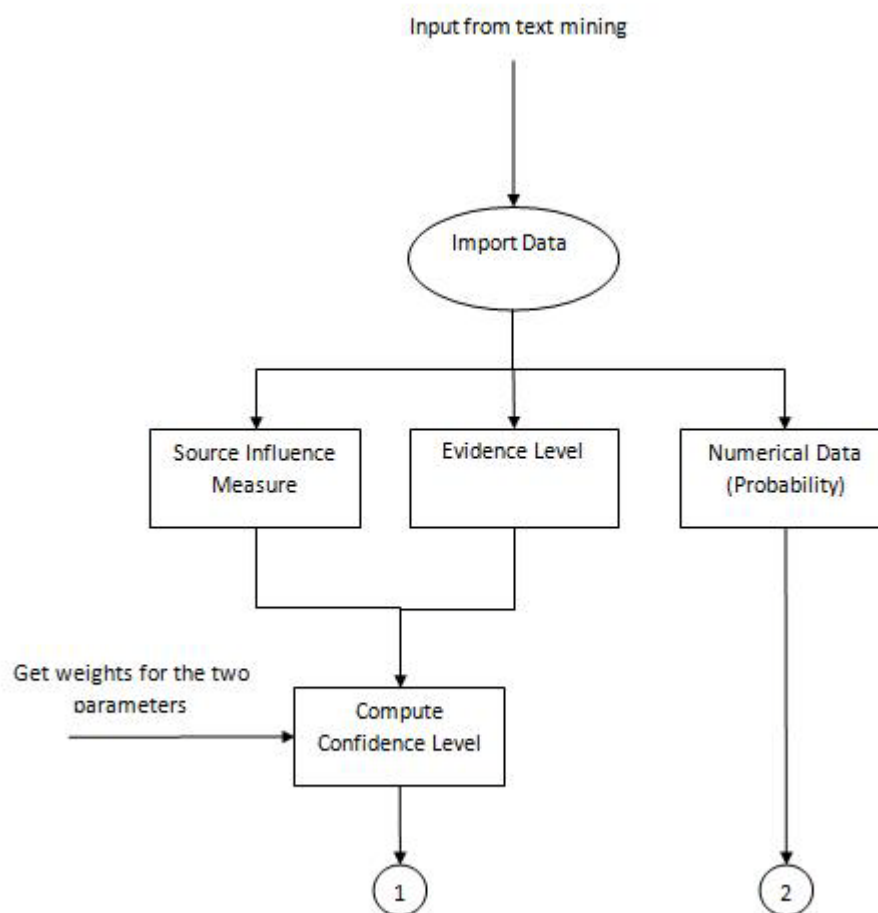
Figure 5.1 Partial Flow Chart for Importing New Evidence and Computing the Confidence Level

CHAPTER 6. INTEGRATING THE DATA WITH THE BAYEISAN NETWORK

6.1. Integration Issues

As mentioned earlier, certain modeling issues need to be resolved while converting causal maps into BNs. As discussed in [4], two most widely used methods are structured interviews and adjacency matrices. In structured interviews, the experts are provided a list of paired concepts as well as different alternative specifications of the relation between the concepts in the original map and asked to choose an alternative to specify the direct relation between the pair of concepts. Using adjacency matrices, the experts are asked to specify for each cell, whether it is a positive, negative or null relation. Though the role of the expert in this process may not be completely eliminated, we can attempt to provide more details to help make the task easier. These details are essentially suggestions for node mapping, loop handling, choosing between direct and indirect relations and values for probabilities in the light of new data.

6.2. Mapping Noun Phrases to Nodes in a Bayesian Network

Mapping the mined noun phrases to a node in the existing BN is a semantic classification problem and can be solved using one of the existing information retrieval and/or classification techniques.

6.2.1. k-nearest Neighbor

Using k-nearest neighbor (k-nn) technique, the new noun phrase can be searched in a space containing all the node names. The Microsoft Full-Text

engine is one such application which can query a search string and return the search result sorted by relevance ranking [21].

6.2.2. Vector Mapping

Another method involves use of vector representation of the names of the nodes in the BN. The new noun phrases are also converted into a vector and compared to all the existing vectors to find a match. These techniques however fail to map semantically equivalent noun phrases.

6.2.3. Machine Learning

For a domain which has a large training data, machine learning techniques such as Weight-normalized Complement Naïve Bayes (WCNB) [22] can be used. The training data consists of a large corpus of semantically mapped noun phrases. This is used by the WCNB algorithm to calculate the prior probability maximum likelihood estimate for every combination of noun in the domain and noun phrase representing a node. This prior probability is then stored in a mapping table which contains a unique row for every combination of noun phrase and node in the domain, as shown in Table 6.1. The noun phrases can be stored in a stemmed format for use by the algorithm. Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form – generally a written word form [27]. Once the training is complete, mapping a noun phrase from text mining to a node in the BN is a simple table lookup to compute the probability of a match. If the probability is above a pre-defined threshold, then a match is deemed to be found.

Table 6.1 Stem Code to Node Mapping Table

Node	Noun	Match Probability
Visual Impairment	vision	0.9124
Visual Impairment	visual	0.9487
Visual Impairment	eyesight	0.8461
Visual Impairment	surrounding	0.0042
Visual Impairment	environment	0.0627
Environmental Hazard	vision	0.0828
Environmental Hazard	visual	0.0285
Environmental Hazard	eyesight	0.0076
Environmental Hazard	surrounding	0.7857
Environmental Hazard	environment	0.8875

6.2.4. New Association

If a match is not found for one or both of the noun phrases with any of the existing nodes, then it means that the association uncovered is not seen before. In this case, the node(s) along with an associating link will have to be created and the mined probability and confidence will be directly assigned to the new association.

6.3. Handling Cycles

The causal association mined could introduce loops in the BN. This needs to be detected and resolved. As discussed in [4], causal loops can exist for two reasons. First, they may be coding mistakes that need to be corrected. Second, they may represent dynamic relations between variables across multiple time frames. While an expert is required to resolve these loops, an automated system can attempt to look at the chronological order of the nodes in the BN. Since the BNs are built from causal maps, they have an implicit chronological order: the cause has to occur before the effect. Any new association, which draws a relation from a node later on in the existing chronological order to a node earlier, can be flagged as either representing a dynamic relationship or a possible error.

As shown in Figure 6.1, the discovery of evidence supporting a causal relation from x3 to x1 will induce a loop in the network and needs to be resolved by a human reviewer. If the new evidence has a significantly lower confidence measure, than the existing links, then it can be discarded as an error. Else, it is possible that the two nodes are interacting across different state levels and might require replicating the network to represent different state of the nodes at different time instances. Then, a link can be created across the nodes in two different networks to represent the state transition.
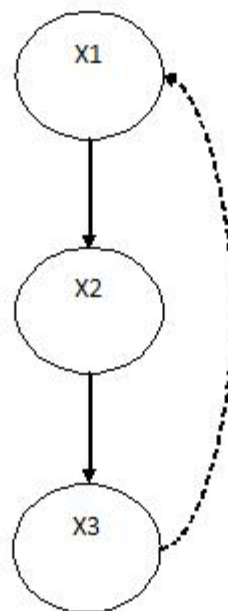
Figure 6.1 Preventing Cycles in the Bayesian Network

6.4. Direct and Indirect Relations

When faced with multiple paths between nodes, as shown in Figure 6.2, the confidence measure can be used as a parameter to decide which path to retain. For each of the path, the average confidence measure over all the edges in the

path can be computed. The path which has the higher confidence measure can be suggested for retaining.
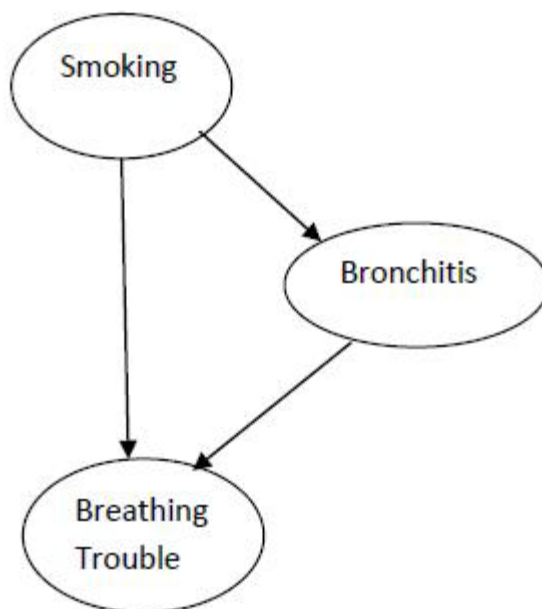
Figure 6.2 Direct and Indirect Relations

6.5. Deriving the Probability

By this stage, the target cause-effect nodes and their corresponding link have either been identified or created new. The next step is to derive the probability for the association and update the conditional probability table for the effect node in the network. There are two cases possible here:

- The association under consideration is accompanied by a probability value.
- No probability value was available for the association.

The following sections discuss these two cases.

6.5.1. Truth Maintenance

For causal relations mined without a probability value, the number of data evidences discovered to support a particular relation can be stored and the probability updated via truth maintenance [25]. In a nutshell, truth maintenance works like this:

To begin with, the belief in a node is equal among all its possible states. If a node is dual stated, like Yes-No, then the belief in each of the two states is 0.5.
As new evidence becomes available, this belief is updated. If an evidence is provided for the state of the node to be "Yes", then the probability of "Yes" changes to 2/3 and that of "No" changes to 1/3.
Similarly, if more evidences are provided for the state of the node to be "Yes", then its probability continually changes to ¾, 4/5, 5/6, etc until it becomes a near certainty.

A truth maintenance system (TMS) maintains consistency between old believed knowledge and current believed knowledge in the knowledge base (KB) through revision [28]. It maintains a record of the current belief set which include supporting evidences and contradictions. It provides the entire belief set to the Inference Engine to make the decisions.

6.5.2. Averaging

For relations mined with a probability value, the prior probability should include the new evidence and also all of the evidences discovered until then. A simple way would be to take an average of all the probabilities. However, the evidences accumulated so far could be from various different sources and could be a result of a wide range of surveys, experiments or pure hypothesis. As discussed before, the confidence measure associated with each of these evidences, attempts to quantify the source of this information. This confidence can now be

used as a weight in calculating the final probability that can be associated with the relation.

Another way to arrive at the new prior probability is to compute a weighted average of the probabilities from all evidences till date, including the newly found evidence, where the weights are the confidence measures, as shown in Eq. 6.1:

$$\text{new\_probability} = \frac{\sum_{i=1}^{n}\left(\text{confidence}_i * \text{probability}_i\right)}{\sum_{i=1}^{n}\left(\text{confidence}_i\right)}$$ Eq. 6.1

$$\text{new\_confidence} = \frac{\sum_{i=1}^{n}\text{confidence}_i}{n}$$ Eq. 6.2

The confidence associated with this association will be an average of the confidence of all the evidences, as shown in Eq. 6.2. This new confidence can now be presented to the expert reviewing the results, to help resolve key issues in updating the Bayesian Network. The new probability and confidence measures replace the existing ones for the association in the network.

There are cases when it is desirable to 'fade' the prominence associated with evidence over time. In other words, as evidence gets older, the confidence associated with it reduces. For such cases, the above equations can be modified to factor in the age of the evidence, by introducing a new parameter which is a function of time as shown in the equations below.

$$\text{new\_probability} = \frac{\sum_{i=1}^{n}\left(\text{total\_confidence}_i * \text{probability}_i\right)}{\sum_{i=1}^{n}\left(\text{total\_confidence}_i\right)}$$ Eq. 6.3

$$new\_confidence = \frac{\sum_{i=1}^{n} total\_confidence_i}{n} \qquad \text{Eq. 6.4}$$

$$\text{where } total\_confidence_i = \left(confidence_i * age\_factor_i\right) \qquad \text{Eq. 6.5}$$

6.6. Identifying the States of the Nodes

The next step in the process is to identify the state space for the nodes. This is an expert driven activity and is not addressed in the thesis. This work assumes a default 'Yes' and 'No' state for every node related by the causal association and the automated updating of CPT in the system developed, occurs only if the nodes are dual stated. However, the methods described in here can be enhanced to mine adjectives which qualify the nouns in question. These adjectives can then be part of the suggestion along with the probability and confidence.

6.7. Resolving Noisy-OR and Noisy-AND

The last step of the process is resolving Noisy-OR and Noisy-AND conditions in the network. This process is not a candidate for automation and requires an expert's knowledge for resolution.

CHAPTER 7. EVALUATION

7.1. <u>The Setup</u>

The system supporting the ideas presented in this thesis was developed as a part of the SCANS system developed by My Health Care Manager. The system, developed as a prototype, demonstrates the processing of information mined from Geriatric health literature. It was built using the .Net framework. The user interface was built using C# and the backend was Microsoft SQL Server 2009 Express edition. The Netica APIs were used to implement Bayesian Network. Netica is a commercial system developed by Norsys Inc.

The core logic is implemented as a library used by the user interface. A part of the logic is also implemented as SQL stored procedures which are executed by the user interface. The backend comprises of a relational database and the stored procedures. Detailed information about the structure of the software and the database is provided in the appendix.

7.2. <u>The Data</u>

The system uses the Influence measures for publications in Geriatric care literature and research articles from these publications for the triplets and probabilities representing the mined information. The Influence measures are available at the ISI Web of Knowledge website [29].

7.3. <u>Software Features</u>

The primary goal of the system is to process mined data and integrate it with existing Bayesian Networks. Certain information can be integrated automatically while the rest are processed in a semi-automated manner. Therefore, an important feature is to provide information to the user for cases when the integration cannot be done in a fully automated way. This system does not contain a text mining utility as yet and the data required from this operation is manually filled into a relational database. As a result, certain additional functionalities are implemented to make the prototype usable. The operations implemented in this software system are:

7.3.1. Normalizing Influence Measure

The influence measures supported by the system are:

- Impact Factor
- Eigen Factor
- Article Influence Score

These scores, in their original form, don't have any specific upper or lower bounds. However, the absolute values of these measures don't have much meaning with respect to our system, unless they are placed in perspective by bounds. The highest and lowest values among these measures can serve as bounds. If desired, the user can specify these bounds as the 'min' and 'max' values. These bounds are then used for normalizing these influence measures.

7.3.2. Importing New Evidence

This operation interfaces text mining with the system. It works on the raw data provided by a text mining utility and prepares it for use by the rest of the system. The steps carried out in this process are as follows:

- Get the weights for Influence measure and Evidence Level assigned by the user. These are the two parameters used for computing the confidence level of the evidence and the default weights assigned by the system are 50% each.
- Get the Influence measure chosen by the user.
- For every evidence:
  - ➢ Get the weight associated with its evidence level and the value of the Influence measure associated with the publication.
  - ➢ Using these, compute the confidence level for the evidence.
  - ➢ Check if the noun phrases have an entry in the database. If it's a phrase not seen before, create a new entry. This step is essential in mapping the nouns to nodes in the Bayesian network.
  - ➢ Check if the nouns have been related by a causal association. If not, create a new relation in the database to represent this new causal association.
  - ➢ Finalize the import by adding the evidence into a table along with its probability and confidence measure.

Figure 7.1 shows the process described above. The outputs of this process, along with those from Figure 5.1, form the new evidence.
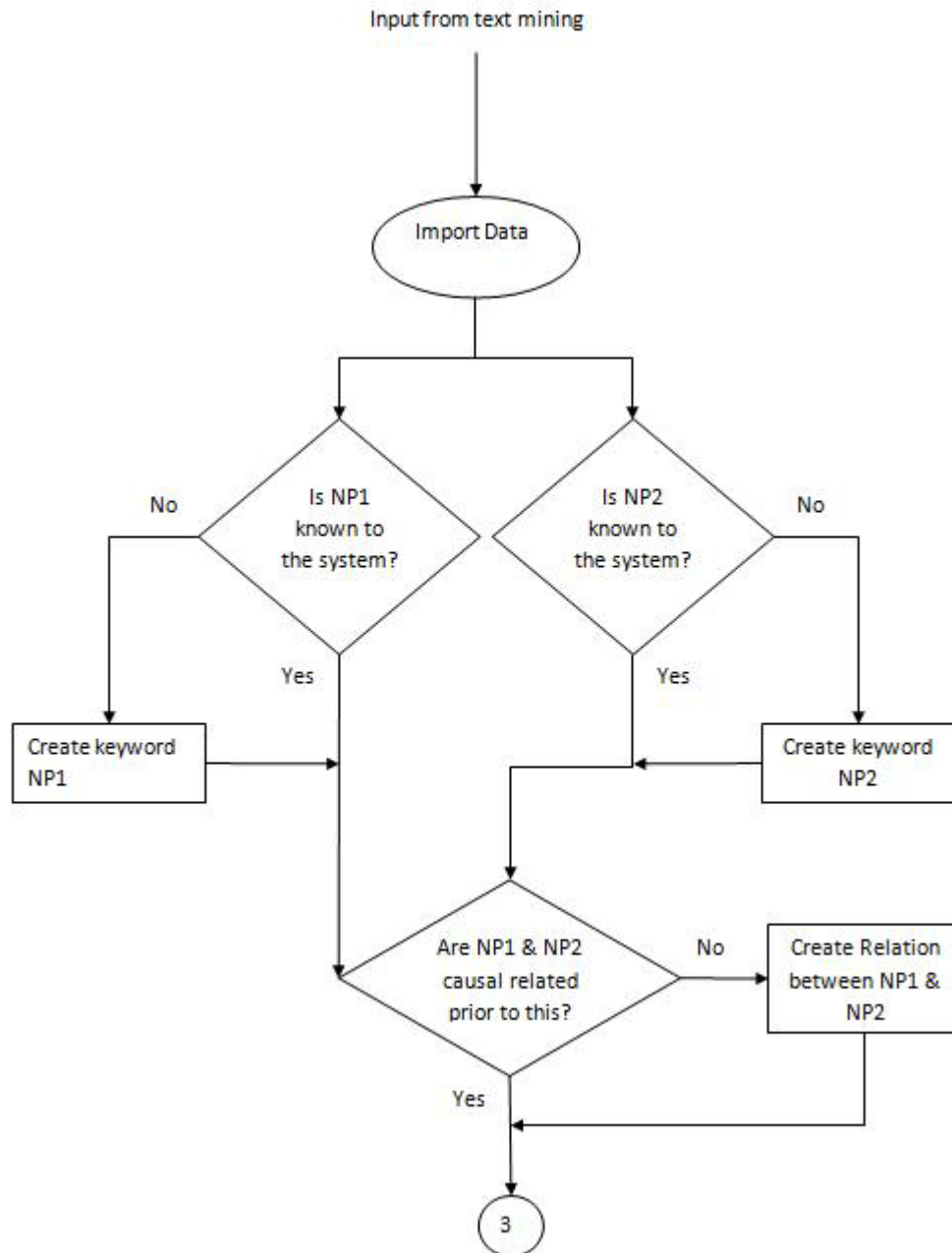
Figure 7.1 Partial Flow Chart for Importing New Evidence from Text Mining into the System

7.3.3. Mapping Nodes to Keywords

This system provides a very rudimentary utility to manually map keywords to nodes. A complete machine learning system which can automatically map

keywords to nodes is complex to implement and logistically not possible within the timeframe of this research work.

The utility displays a column with the available keywords in the system and another column with unmapped keywords as shown in. The user can select the keyword and the matching node to create a mapping. This also helps in resolving synonyms because all different representations of the noun phrase get mapped to the same node. Since the system being developed is still at a prototyping stage, this method of manually mapping is sufficient. But a full scale implementation will have to address this task which can potentially be a major bottleneck in the operation.

7.3.4. Generating Suggestions

This step consolidates all the evidences and writes out the result into a database table. This step identifies all the unique triplets based on the nodes mapped to them and computes the probability and confidence based on the equations Eq. 1 and Eq. 2 discussed earlier. The nodes representing cause-effect relation are also written out with the result. If the evidence is new and has no associated representation in the Bayesian Networks, then the triplet along with its probability and confidence is written out as it is but the fields representing the cause-effect nodes are left 'null' to indicate that it's a new causal association. The generated suggestions are then displayed on the screen for review by the user.

For causal associations already existing in the BN, the previous probability and confidence is displayed to facilitate comparison with the newer values.
For causal associations which induce loops in the BN, a message is displayed indicating the same.

7.3.5. Reviewing Suggestions

Once the suggestions are generated and displayed on the screen, the user can review them by selecting the interesting suggestion and clicking 'review' button. The system then attempts to display this suggestion as part of the appropriate Bayesian Network:

Case1: Both the nodes and the link between them exists. In this case, only the conditional probability table needs to be updated. The system has been designed to handle Boolean states of the nodes, i.e. the two possible states of the nodes are 'Yes' and 'No'. This means that for a causal relation like "Falls lead to death 20% of the time" the corresponding CPT representation would be modified as shown in Table 7.1. A similar example from the system is shown in Figure 7.2 and Figure 7.3 below.

Table 7.1 Modified CPT Representation at Node 'Death'

Parent1	Fall	No	Yes
No	No	…	…
No	Yes	0.8	0.2
Yes	No	…	…
Yes	Yes	…	…

Figure 7.2 Case 1: Evidence to be Reviewed

Figure 7.3 Case 1: Updating the CPT

Case 2: Both nodes exist in the network but are not linked causally. Create the link if it does not induce any loops in the network. As shown below, the new evidence linking the client's gender and arthritis is applied to the BN. Figure 7.4 shows the original BN and Figure 7.5 shows the BN after applying the new evidence.
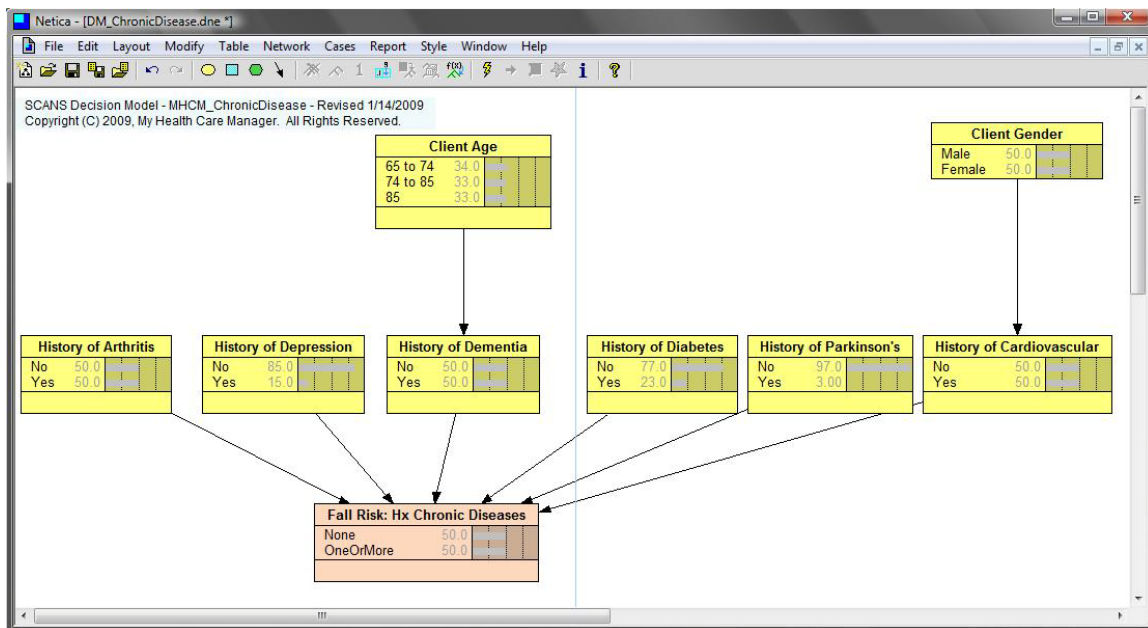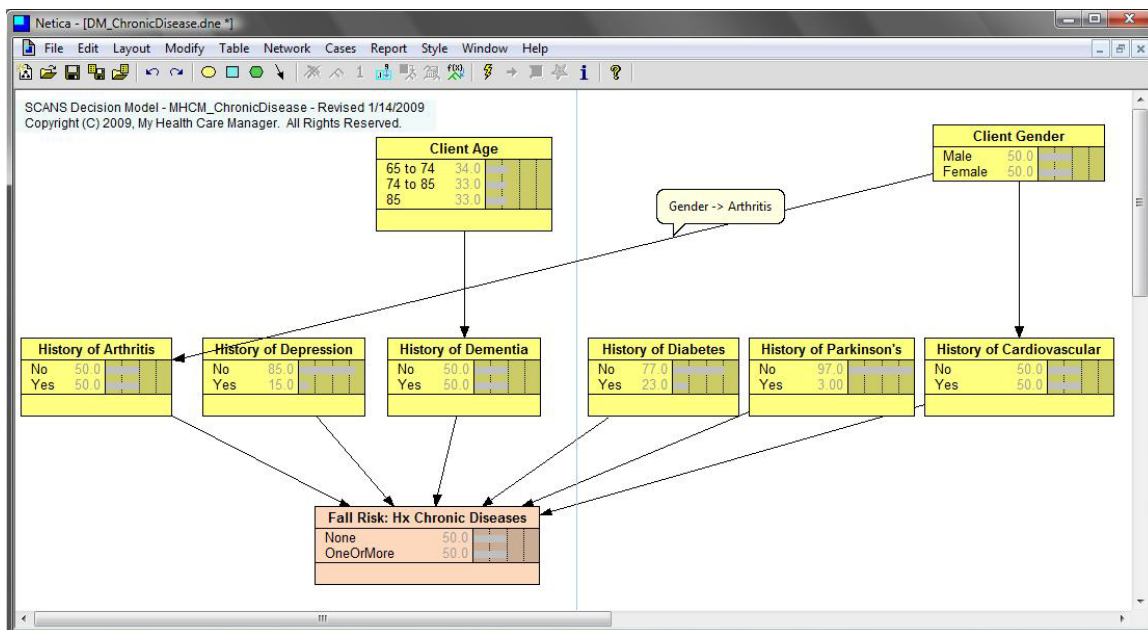
Figure 7.4 Case 2: BN Before Updating with New Evidence

Figure 7.5 Case 2: BN After Adding the New Link

Case 3: Special case of case 2. Both nodes exist but the directionality of the new causal association is reverse of what currently exists. Loop detection utility checks for this condition and recommends the link with higher confidence. This might lead to the existing link being replaced by the newer link.

Case 4: One of the noun phrases in the causal relation has a corresponding node while the other doesn't exist. The second node needs to be created along with the link between the nodes. If a 'cause' node was created, then the CPT at the 'effect' node needs to be updated. This usually results in an exponential increase in the number of rows in the CPT to accommodate all the combinations arising out of the states in the new node. Figure 7.6 shows the original BN before adding the new evidence. Figure 7.7 shows the evidence to be reviewed. Figure 7.8 shows the BN with the new evidence incorporated in the form of a new cause node and the Figure 7.9 shows the CPT updated the effect node.

If the new node is an 'effect' node, then it probably is an intermediate node. If it turns out to be a leaf node of an existing BN, it could probably suggest a new intervention. If it not a leaf node, then it would result in a dangling intermediate node which has no impact on reasoning and needs to be re-wired into the existing network by an expert.
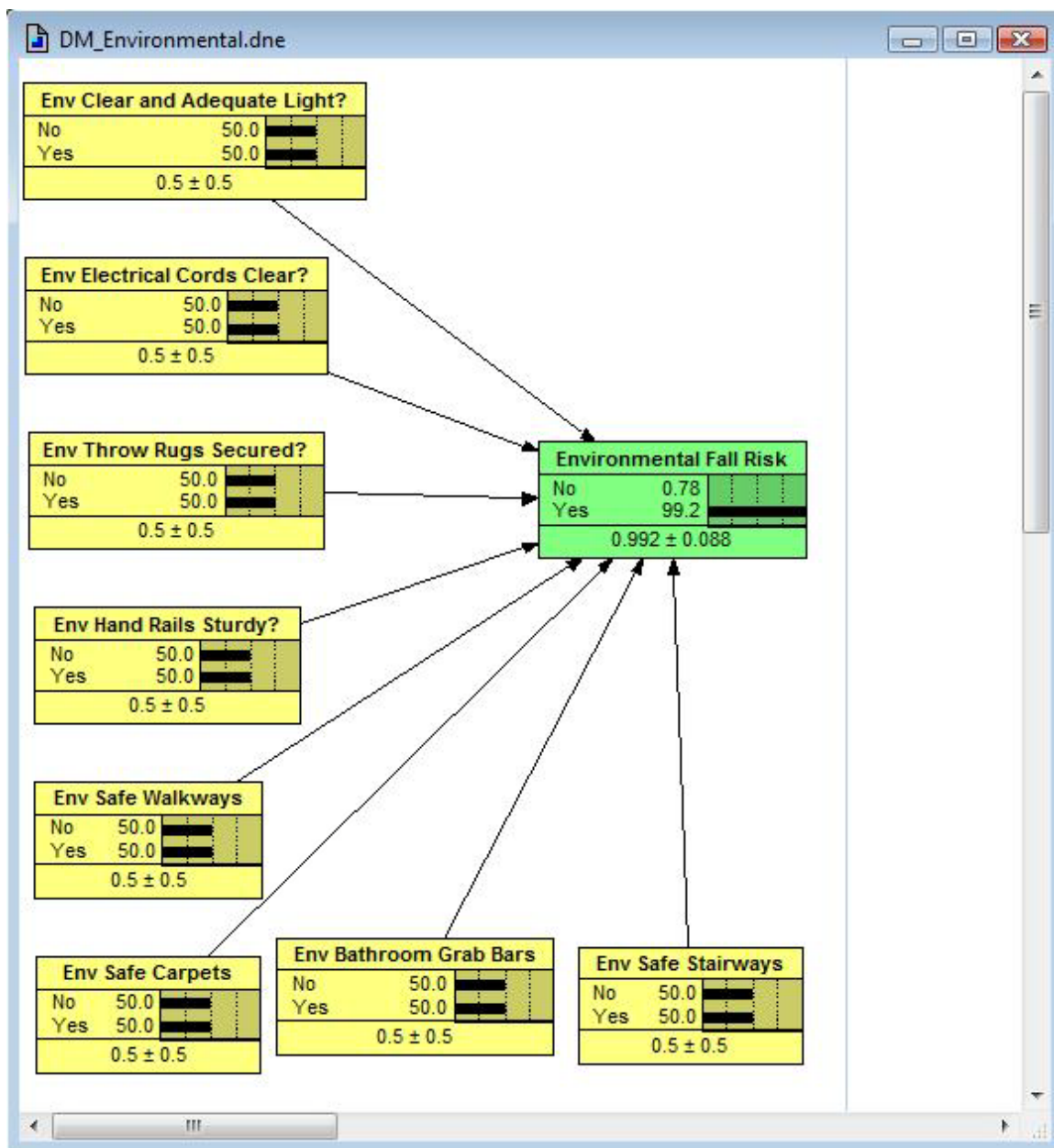
Figure 7.6 Case 4: Original BN

Chosen for Review	Suggestion_ID	Cause_Keyword	Effect_Keyword	Source_Node	Target_Node	Probability	Confidence
☐	201			Env Safe Stairways	Environmental Fa...	0.74824	0.71790
☐	202			Env Electrical Co...	Environmental Fa...	0.44620	0.80360
☐	203			Env Safe Walkw...	Environmental Fa...	0.33659	0.77583
☐	204			Env Clear and Ad...	Environmental Fa...	0.37848	0.82555
☐	205			Client Gender	History of Arthritis	0.90000	0.72500
☐	206	rugs and mats	fall		Environmental Fa...	0.69606	0.68875
☑	207	obstacles	fall		Environmental Fa...	0.63091	0.81305
☐	208	stepovers	fall		Environmental Fa...	0.55936	0.83805
☐	209	wet bathroom floor	fall		Environmental Fa...	1.00000	0.74310
☐	210	obesity	arthritis		History of Arthritis	0.80000	0.65000
☐							

Figure 7.7 Case 4: Evidence to be Reviewed
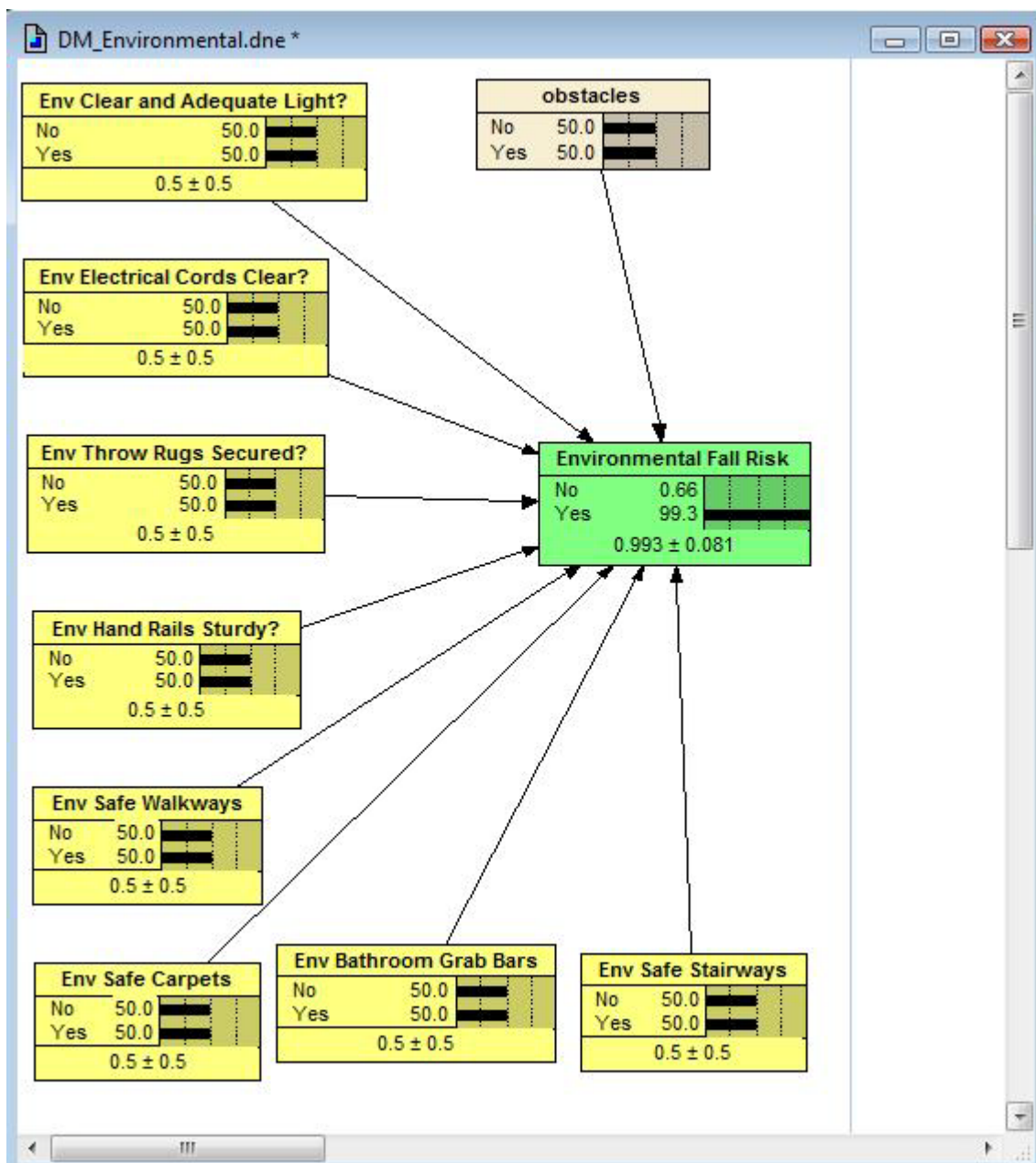
Figure 7.8 Case 4: BN After Adding the New Evidence

Figure 7.9 Case 4: CPT Updated at Node 'EnvFallRisk' After Adding New Cause Node 'obstacles'

Case 5: The causal relation has never been seen before and both the nodes and the link need to be created. This case most likely leads to the beginning of a new Bayesian network.

CHAPTER 8. CONCLUSION

8.1. <u>So Far</u>

This thesis studies the problem of updating a Bayesian Network from literature. It breaks down the problem and proposes a step by step approach to solving it. The thesis studies existing techniques which can be used in solving some of the steps and proposes some new techniques for the rest. It builds on the technique of mining causal associations from text and using the resulting causal associations to update the Bayesian Network. It proposes the use of an influence measure for the source journal and an evidence level for the causal evidence mined. It proposes to use confidence measure as an instrument in evaluating new evidence mined from text and presents techniques to derive a confidence measure for causal associations mined. It further presents ways to partially automate resolution of the modeling issues by providing the expert with meaningful alternatives computed using the confidence measure of the edges in the BN. The thesis also identifies gaps in research and techniques that need to be developed to facilitate a more complete system to mine data from literature which is relevant to updating a Bayesian Network.

8.2. <u>Future Work</u>

Future work will focus on medical domain since higher occurrence of causal patterns was found in it, given that diseases can be diagnosed or cured by recognizing their causes as well as the effects of prescriptions. Moreover, future work will include improved evaluation methods and term extraction methods. Along with more focused evaluations, effort needs to be put into measure how well the network works when performing tasks such as obtaining accurate

inferences, answering questions about the content of the text or supporting decision-making. For term identification, ontologies formed by modifiers and nouns can be considered, the recognition of specialized terms of the topic when generalization takes place as well as the integration of anaphora resolution. Other specific areas of research identified by this work are:

- ➢ Identifying the degree of causality encoded in the text through the usage of auxiliaries (as may, could and must) as well as adverbs (such as strongly, slightly) so that more precise probabilities can be derived.

- ➢ Use of ontologies can be researched with respect to a BN. One definite advantage know at this point is the mapping of completely new information to an existing BN. The current system does not have the capability to map a causal association to a BN if both the nodes are new. By maintaining a mapping of BN and its nodes with an ontology, new causal associations can be automatically mapped to an existing BN, by semantically classifying the source of the evidence to a known term in the ontology.

- ➢ Identifying state of the nodes of the BN by using adjectives associated with the nodes in the causal association.

- ➢ Using inhibitors to derive interventions for the BN. For example: "Vaccination prevents flu". 'Vaccination' can be used as an intervention for 'Flu'.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] K. Anthony, "Introduction to Causal Modeling, Bayesian Theory and Major Bayesian Modeling Tools for the Intelligence Analyst", USAF National Air and Space Intelligence Center (NASIC).

[2] R. Girju and D. Moldovan, "Text Mining for Causal Relations", Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference, pp. 360-364, 2002.

[3] O. Sanchez-Graillet and M. Poesio, "Acquiring Bayesian Networks from Text", Proceedings of LREC, Lisbon, May 2004.

[4] S. Nadkarni and P. Shenoy, "A Causal Mapping Approach to Constructing Bayesian Networks", Decision Support Systems 38, pp. 259-281, 2004.

[5] J. Pearl, "Bayesianism and Causality, Or, Why I Am Only A Half-Bayesian", in Foundations of Bayesianism, Kluwer, 2001.

[6] J. Pollock, "Causal Probability", Synthese, volume 132, pp. 143-185, Jul 2002.

[7] HTTP://EN.WIKIPEDIA.ORG/WIKI/CAUSALITY, Oct 2009.

[8] HTTP://EN.WIKIPEDIA.ORG/WIKI/CAUSAL_MODEL, Oct 2009.

[9] D. Heckerman, "Bayesian networks for Data Mining", Data Mining and Knowledge Discovery, 1:79-119, 1997.

[10] HTTP://EN.WIKIPEDIA.ORG/WIKI/EVIDENCE-BASED_MEDICINE, Oct 2009.

[11] J. Pearl, "Causality", Cambridge University Press, New York, 2000.

[12] HTTP://EN.WIKIPEDIA.ORG/WIKI/IMPACT_FACTOR, Sep 2009.

[13]   N. Friedman, M. Goldszmidt, "Learning Bayesian Networks with Local Structure". Proceedings of the Twelfth Conference for Uncertainty in Artificial Intelligence (UAI-96), pp. 252-262, San Francisco, CA. Morgan Kaufmann Publishers, 1996.

[14]   J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Mateo, CA, 1988.

[15]   S. Nadkarni and P. P. Shenoy, A Bayesian network approach to making inferences in causal maps, European Journal of Operational Research, 128, No. 3, 2001.

[16]   C. S. Spetzler and C. S. Staël von Holstein, Probability encoding in decision analysis, Management Science, 22, No. 3, 1975.

[17]   HTTP://EIGENFACTOR.ORG/, Oct 2009.

[18]   HTTP://WWW.THOMSONREUTERS.COM/PRODUCTS_SERVICES/SCIENTIFIC/JOURNAL_CITATION_REPORTS, Oct 2009.

[19]   HTTP://EN.WIKIPEDIA.ORG/WIKI/PAGERANK, Oct 2009.

[20]   HTTP://EN.WIKIPEDIA.ORG/WIKI/EIGENFACTOR, Sep 2009

[21]   HTTP://MSDN.MICROSOFT.COM/EN-US/LIBRARY/MS142587.ASPX, Oct 2009.

[22]   J. Rennie, L. Shih, J. Teevan, D. Karger, "Tackling the Poor Assumptions of Naïve Bayes Text Classifiers". In Proceedings of the Twentieth International Conference on Machine Learning, 2003.

[23]   HTTP://WWW.CONSULTGERIRN.ORG/EVIDENCE-BASED_PRACTICE, Oct 2009.

[24]   O. Sanchez-Graillet, M. Poesio, "Acquiring Bayesian Networks from Text", Proc. of the Fourth LREC, Lisbon, May 2004.

[25]   "In Praise of Bayes", Economist 356, no. 8190, 30 Sep 2000: 83-84.

[26]   A.W. Rader, V.M. Sloutsky, "Conjunction bias in memory representations of logical connectives", Memory & Cognition, 29(6), pp. 838-849, 2001.

[27]   HTTP://EN.WIKIPEDIA.ORG/WIKI/STEMMING, Sep 2009.

[28]   HTTP://EN.WIKIPEDIA.ORG/WIKI/TRUTH_MAINTENANCE_SYSTEM,
         Sep 2009.

[29]   HTTP://ADMIN-APPS.ISIKNOWLEDGE.COM/JCR/JCR, Oct 2009.

APPENDIX

APPENDIX

This section describes the software system built to demonstrate the ideas presented in this thesis. The ER diagram for the backend relational database is shown in Figure A.1.
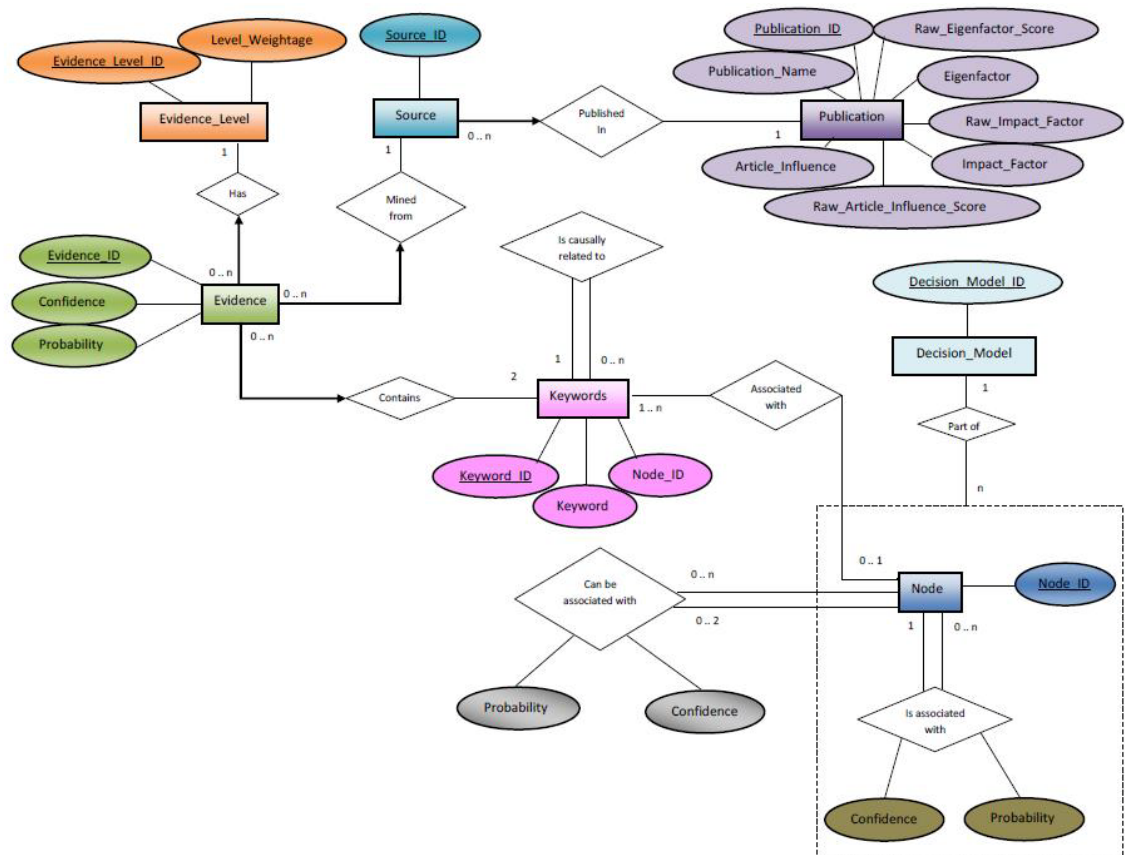

Figure A.1 ER Diagram for the Relational Database Schema

Table A.1 Raw_Evidence

Triplet_ID	Source_ID	NP1	NP2	Verb	Probability	Evidence_Level_ID
2	15	steps	fall	23	0.70000	2

The database schema is briefly described in this section. The table Raw_Evidence contains the output of text mining. It contains the causal association in the triplet form, along with the probability, source of the evidence and the evidence level.

Table A.2 Publication

Name	Raw Impact Factor	Impact Factor	Raw Eigenfactor Score	Eigenfactor	Raw Article Influence Score	Article Influence
AGE	2.9250	0.43551	0.00256	0.05320	0.0000	0.00000

The table Publication contains the list of publications which are the primary source of literature. This table contains the various influence measures used by the system for every publication. The 'Raw' fields are publically available data while the others are normalized values based on user defined minimum and maximum. This table does not change very frequently. For example, it needs to be updated only when the influence measures shown above are updated annually or when a new publication is used.

Table A.3 Evidence_Level

Evidence_Level_ID	Level_weightage	description
1	1.0000	Level1: Systematic Reviews
2	0.9500	Level2: Single Experimental Studies
3	0.9000	Level3: Quasi Experimental Studies
4	0.8500	Level4: Non-Experimental
5	0.8000	Level5: Case Report, Program Evaluation
6	0.7500	Level6: Opinion of Respected Authorities

Table Evidence_Level is a static table containing the definition for every level of evidence used in the system. The primary field here is the weight associated with the level.

Table A.4 Source

Source ID	Publication ID	Date	Title	Author
1	18	2001	Guideline for the Prevention of Falls in Older Persons	AGS Board of Directors

The table Source contains details of individual articles, papers etc. It is bound to the publication where it appeared.

Table A.5 Keywords

Keyword_ID	Keyword	Node_ID
54	steps	9
55	fall	1

The table Keyword contains all the noun phrases known to the system and a mapping to the node they correspond to in the BN. This table also serves to map synonyms to a single node.

Table A.6 Relation

Relation_ID	Cause	Effect
43	54	55

The table Relation represents the causal association between noun phrases. It can potentially contain a combination for every synonym of every known association. If a 'cause' has two synonyms and an 'effect' has two synonyms, then the table can potentially have four different tuples representing all the combinations.

Table A.7 Evidence

Evidence_ID	Relation_id	confidence	probability
2	43	0.7431	0.7000

The table Evidence represents processed data where the causal association from text mining has been codified and the confidence level has been derived. This table store the result of the 'Import' functionality provided by the system and is used to generate suggestions to the user.

Table A.8 Decision_Model

Decision Model_ID	Name	Description	Filename	Evaluation Order
3	DM_Environ mental	Test Model	DM_Environ mental	3

The table Decision_Model contains details of the Bayesian networks in the system and their physical representation in the machine.

Table A.9 Node

Node_ID	Decision Model_ID	Description	Node_Name	Node_Title
1	3	Aggregation of all Environmant al hazards which can cause fall	EnvFallRisk	Environment al Fall Risk

The Node table contains details of the nodes in the Bayesian network and is mapped to the BN it belongs to.

Table A.10 Association

Association_ID	Source Node	Target Node	Probability	Confidence
1	2	1	NULL	NULL
2	3	1	NULL	NULL

The table Association is an equivalent of the table Relation except that it represents the real causal links which exist in the BNs in the system. For every node to node causal link, there is a corresponding entry in this table.

Table A.11 Suggested_Association

Suggestion ID	Evidence ID	Source Node	Target Node	Probability	Confidence
201	10	9	1	0.74824	0.71790

The table Suggested_Association stores the suggestions generated by the system using the evidence provided to it. It contains the final output of the system. The suggestions are stored in this table for review and can be loaded in the display. Before beginning a session with new set of evidences, the user must clear this table.

The following SQL stored procedures have also been in addition to the C# software:

- [Publication_Normalize_Influence_Score]: Used to normalize the influence measures of the publications.
- [Evidence_Compute_Confidence]: Used to compute confidence level for each evidence in the evidence table.
- [Association_Compute_Probability]: Used to aggregate the evidences and populate them into the Suggested_Association table.

The opening screen of the software utility is shown in Figure A.2. It consists of a left panel with the operations supported and a right panel which displays the data

and results of the operation. The right panel contains tabs with grids, one each for Normalize/Import, Map keywords and Suggestions.
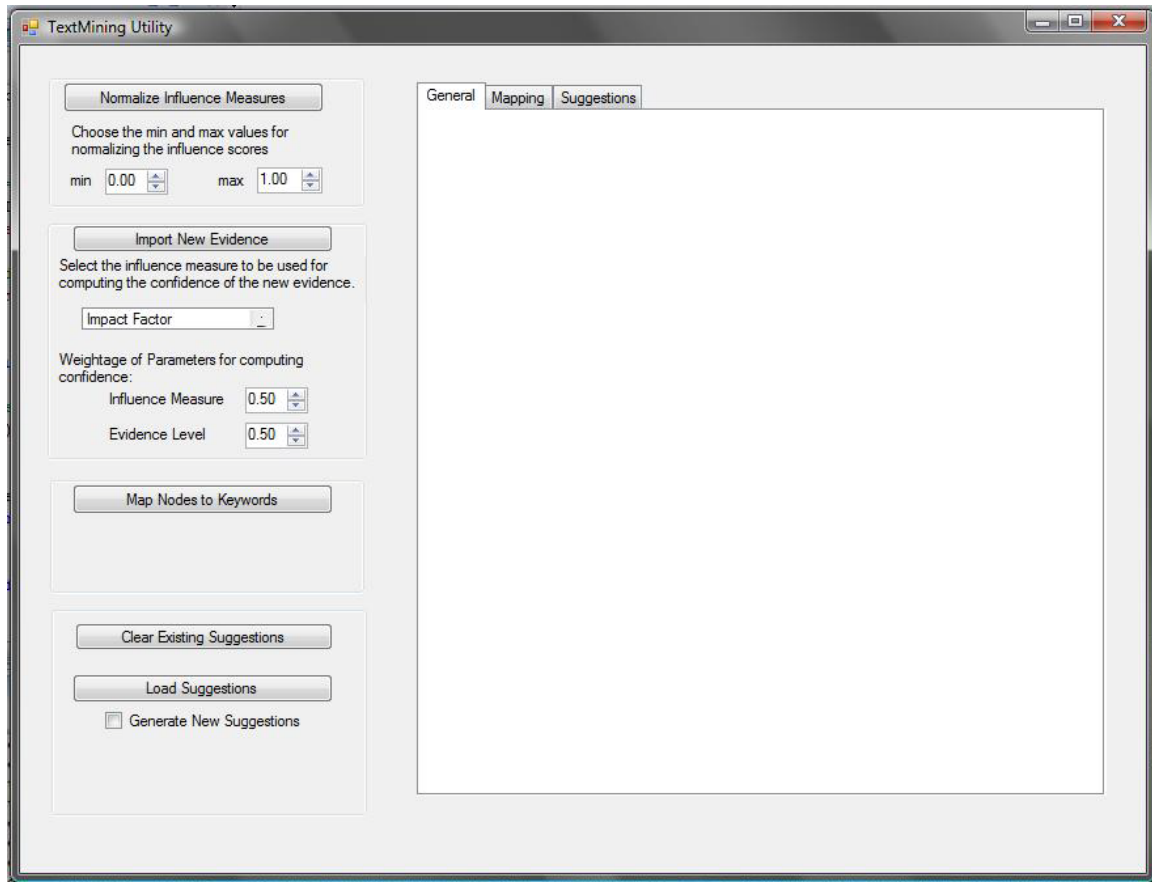
Figure A.2 The Software Utility for Processing Information from Text Mining

The controls for the Normalize Influence measure operation is shown in Figure A.3. The new minimum and maximum values need to be specified by the user and the influence measures will be normalized to that range on clicking the button provided. The values are written out to the corresponding columns in table Publication.

Figure A.3 Normalizing the Influence Measures for the Publications

The controls for Import New Evidence operation are shown in Figure A.4. The influence measure to be used for computing the confidence level of the evidence can be specified in the list provided. The default is Impact Factor. The weights to be used with the influence measure and evidence level can be specified in the numeric up-down control. The default weights are 0.50. The Importing operation is triggered on clicking the button provided and table Evidence is populated. In case the table contains data from previous sessions, a warning message is displayed before beginning the operation. If the user decides to go ahead, then the table is cleared out before beginning the new import.

Figure A.4 Importing New Evidences into the System for Processing

The controls for mapping keywords to nodes are shown in Figure A.5. The operation can be initiated by clicking the button in the left panel upon which two lists are displayed on the panel in the right. These lists contain the list of unmapped keywords and the list of available nodes in the system. The user needs to choose a keyword by clicking on it and choosing a mapping node by clicking on it. The 'submit' buttons provided below the lists, commits the changes to the database.
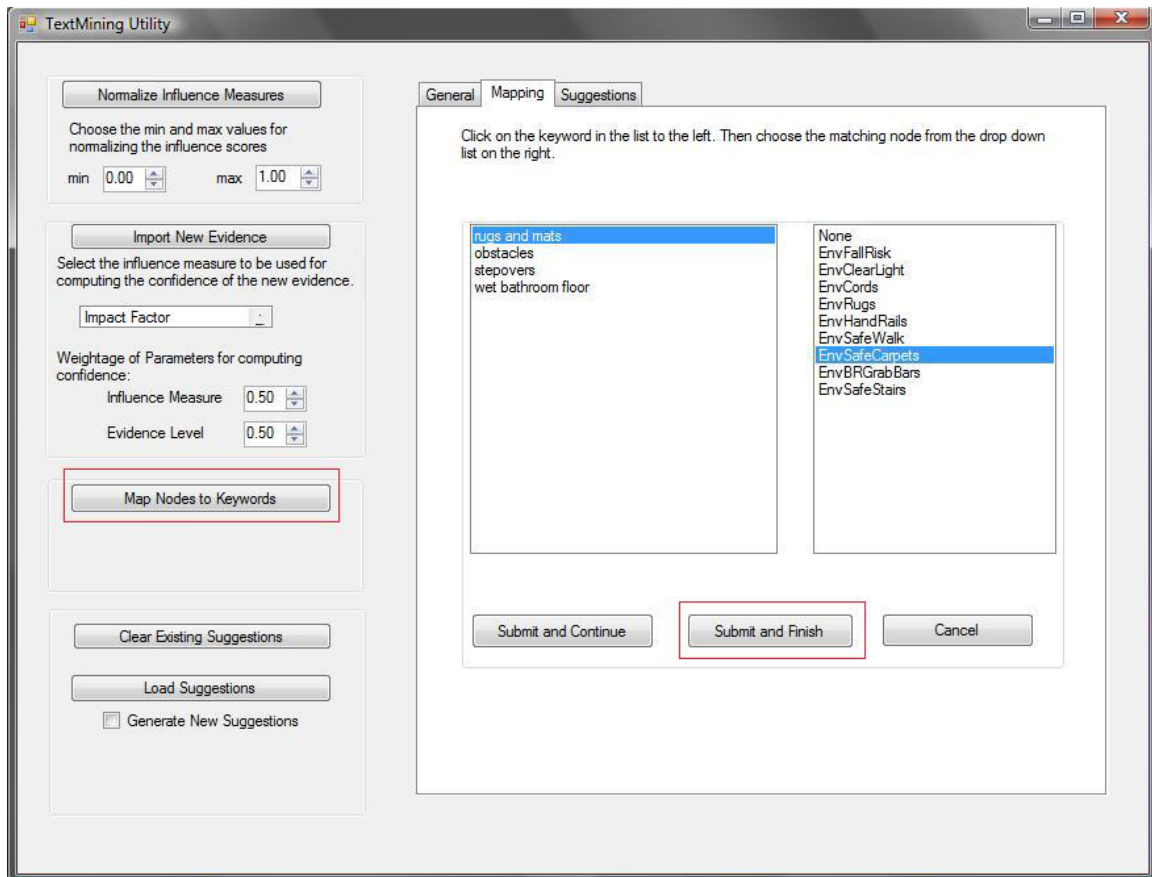
Figure A.5 Mapping Keywords to Nodes in the Bayesian Network

Figure A.6 shows the clearing of table Suggested_Association before generating new suggestions.
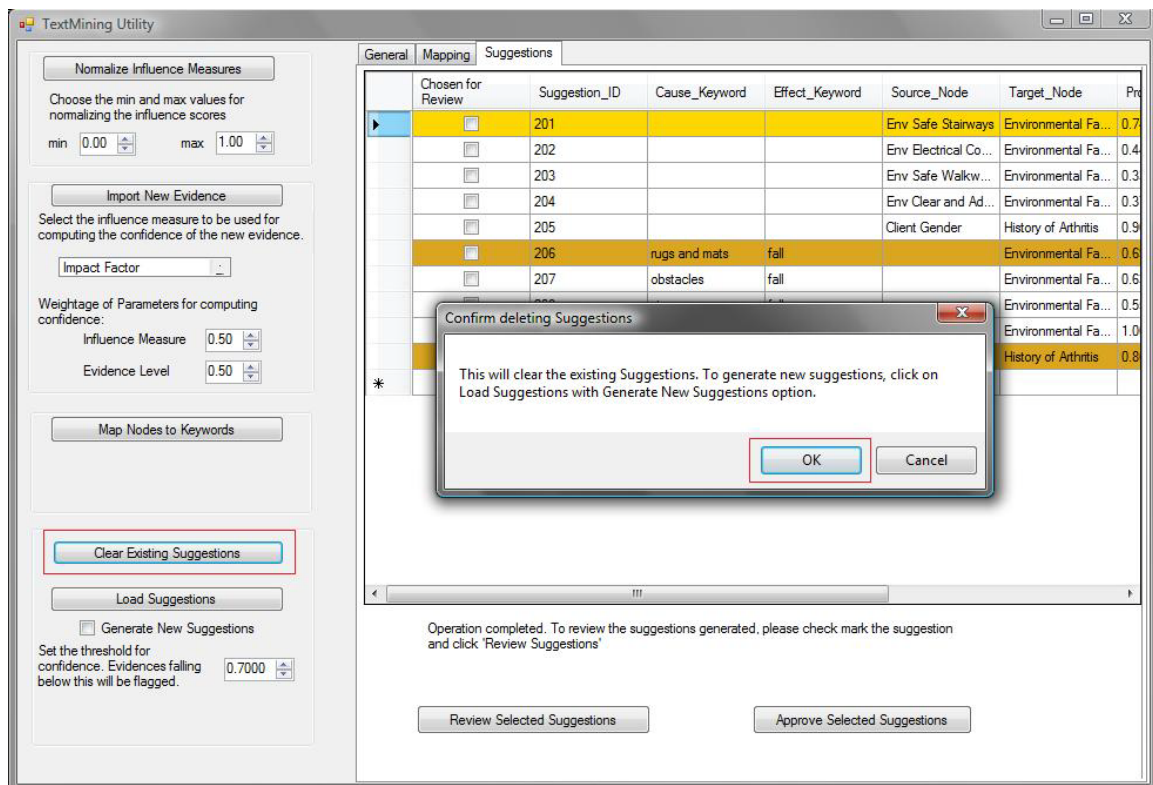
Figure A.6 Clear Suggestions Before Generating New Ones

The last set of controls provided help the user to generate and review suggestions. The Load suggestions button shown in Figure A.7 reads the table Suggested_Association and loads it into a grid display. By checking the Generate new suggestions checkbox, the user can generate fresh suggestions based on the current data in the Evidence table. The grid displaying the suggestions contains a checkbox for selecting evidences of interest. The user can check the evidences and click Review button. The system then attempts to generate a visual representation of the evidence by loading the corresponding BN in Netica. For certain cases, the system also shows some message in the comments column of the grid. This column provides information about the previous confidence level, loops being induced etc.
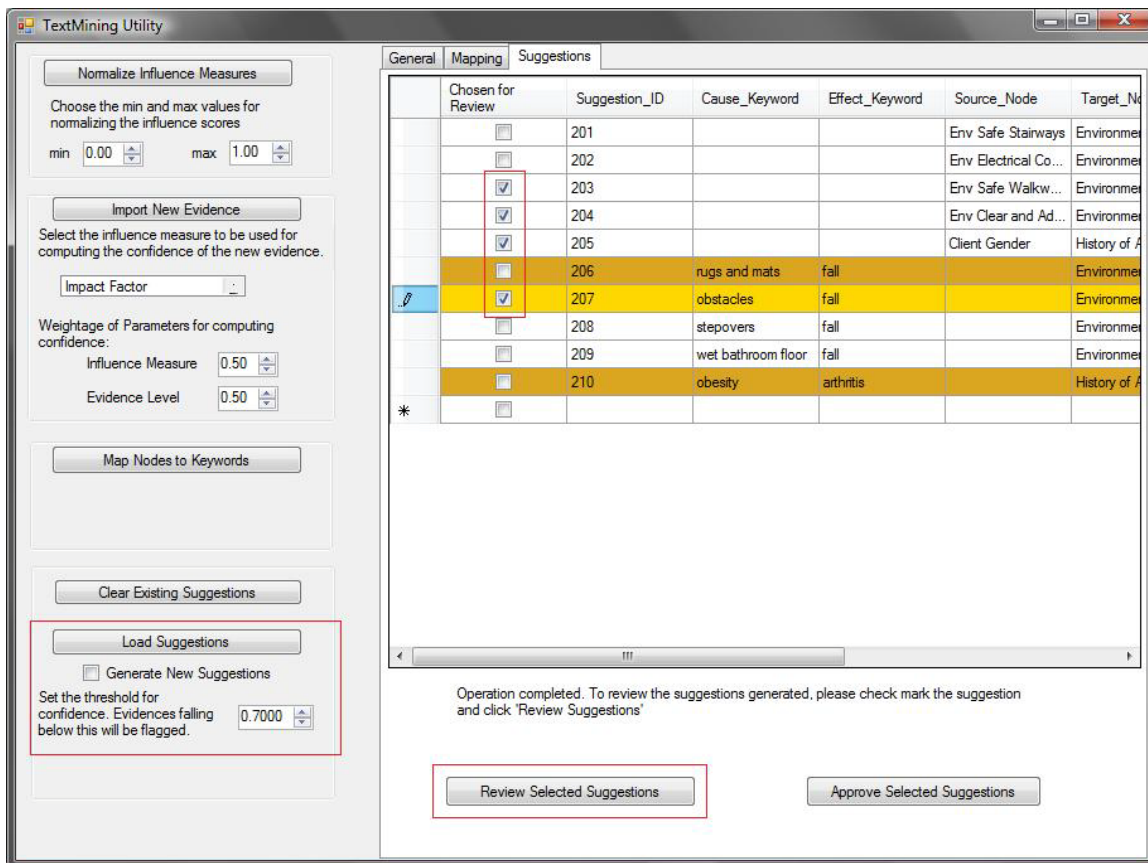
Figure A.7 The Software Utility for Processing Information from Text Mining