AUTOMATIC DETECTION OF ASSOCIATIONS
AMONG TERMS RELATED TO ALZHEIMER’S DISEASE
FROM MEDLINE ABSTRACTS

Dongbing Lai

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Master of Science
in the School of Informatics,
Indiana University

May 2003
Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Master of Science.

Snehasis Mukhopadhyay, Ph.D.

Douglas Perry, Ph.D.

Narayanan Perumal, Ph.D.
# Table of Contents

- Introduction ............................................. 1
- Background ............................................. 9
- Methods ............................................... 17
- Results ............................................... 29
- Conclusion ............................................ 45
- Discussion ........................................... 47
- References ........................................... 49
I. Introduction:

*Brief description of Alzheimer's disease:*

Alzheimer's disease is a progressive, age-related, degenerative brain disorder, which is one of the most serious diseases in old people. The patients' memory is lost and their personality and behavior are changed gradually; furthermore, this process is irreversible until the patients die [1]. Alzheimer's disease first attacks the entorhinal cortex; then to the hippocampus, which help to control short-term memory; then to other regions, especially the cerebral cortex, which is very important in using language and reasoning [1, 2]. After its attack, the neurons degenerate and lose synapses and eventually die [1, 2].

According to the age of having this disease, Alzheimer's disease can be divided to early-onset (usually at age 30 to 60) and late-onset (at age of 65 or older) [1]. About 5% to 10% of Alzheimer's disease cases are early onset [1]. Another way to describe Alzheimer's disease is according to the inheritance pattern. In this way, Alzheimer's disease also can be divided to: sporadic Alzheimer's disease, which has no certain inheritance pattern; and familial Alzheimer's disease (FAD), which has certain inheritance pattern [1]. All FAD are early onset [1].

Alzheimer's disease is a progressive disease and the progression of symptoms can be divided into mild, moderate and severe phases [2, 3]. The symptoms of mild Alzheimer's disease include loss of memory, disorientation, and difficulty of performing routine tasks. [2]. Patients in this phase can live
independently [3]. The moderate symptoms include having great difficulty in daily living, wandering, personality changes, agitation and anxiety [2]. Patients in this phase should be cared by other people. People in severe phase lose all communication functions, almost cannot think, and need total care [2].

**Diagnosis of Alzheimer’s disease**

Early diagnosis is very important to Alzheimer’s disease patients and can delay and slow down the progress. However, the neuritic plaques and neurofibrillary tangles found by autopsy are the only definite markers of Alzheimer's disease [2]. Fortunately, there are lots of methods that can be used to diagnose Alzheimer's disease. Through the modern diagnosis tools, about 90% accuracy can be obtained [2]. Some typical tools include: using Computerized-tomography (CT) or other instrument to do the brain scan; a complete and detailed information about the patient and family's medical history; standard medical test such as blood, urine etc.; neuropsychological test such as memory, language etc. [4].

**Treatment of Alzheimer’s disease:**

Currently there is no treatment that can stop or reverse the Alzheimer's disease. Thus the most important thing for Alzheimer’s disease patients is to control symptoms and slow down the progress [1, 2]. The FDA (Food and Drug Administration) has approved three medications for Alzheimer’s disease: tacrine (Cognex), donepezil hydrochloride (Aricept), and rivastigmine (Exelon) [1]. All of
them act to inhibit the function of acetylcholine [1]. As it is mentioned before, all of these drugs cannot stop the progress of Alzheimer's disease; however, they may slow down the progress of some symptoms for a certain time for those in mild or moderate stages [1]. Currently there are 38 clinical trials in the USA [5].

**Research on the cause of Alzheimer's disease:**

The research on Alzheimer's disease is one of the most active areas in biomedical science. Most of these researches focus on the cause, diagnosis and treatment of Alzheimer's disease. The actual cause of Alzheimer's disease is still not clear. People only know it has a very complex cause, which may involve the interaction of many genetic and environmental factors [1]. Some molecules and changes of their amount and constitution are thought to be related to Alzheimer's disease. For example, in Alzheimer's disease patients, the level of acetylcholine drops dramatically [2]. Other neurotransmitters, such as serotonin, somtostatin and noradrenaline are also considered to be linked to Alzheimer's disease because their concentration in Alzheimer's disease patients is lower than normal [2]. Also, the receptor of these neurotransmitters and all proteins or other molecules in the way of message transmission may play an important role, e.g., some researchers found abnormal phospholipids in Alzheimer's disease patient's neuron, which may influence the function of the receptor [2]. Glucose decreases greatly when neurons degenerate and it is thought to be related to Alzheimer's disease [2]. Studies about this have found capillary, GLUT1, GLUT3, cytochrome oxidase and glutamate may have important role, for example, some studies
showed glutamate and glucocorticoids are related to the increase of calcium level and thus cause the cell die [2].

One important protein related to Alzheimer’s disease is the beta amyloid. It is the component of neuritic plaques and it comes from its precursor protein, which has the name Amyloid-precursor-protein (APP) [2]. Some researchers thought that for some reasons, beta amyloid is insoluble and is toxic to neurons, or it can release free radicals [2]. Others found that it can form channels in neuron membrane and cause the calcium into the cell or disrupt potassium channel, thus can cause the death of the cell [2]. Another important protein is tau. It is the component of neurofibrillary tangles [2]. In Alzheimer’s disease patients’ cells, the tau forms paired helical filaments while in normal cell it is the connector factor between microtubules [2]. There is also accumulation of alpha-synuclein in plaques, thought to interrupt cell-cell communication [1].

Researchers have located some candidate genes/regions on human genome. The mutated APP gene on chromosome 21 has been found in some FAD families. In other FAD families researchers found mutated presenilin 1 on chromosome 14 and presenilin 2 on chromosome 1 [1]. The presenilins may be the enzymes that clip APP and are related to cell death and maintaining cell-cell connection [1]. All these genes are autosomal dominant inheritance [1]. The most important gene related to Alzheimer’s disease is ApoE4. Some studies showed that ApoE4 can bind beta amyloid and make it insoluble, while other found ApoE4 could make the structure stabilized by tau unstable [2]. Also some researchers found that patients with ApoE4 have shorter dendrites [2].
Population-based studies also found one region on chromosome 12 thought to be linked to Alzheimer's disease; the candidate gene is alpha-2 macroglobulin [1]. The genes for pro-inflammatory agents, such as interleukin 1(IL-1) are also found related to Alzheimer's disease [1].

Some environmental factors also have important effects. Too little zinc maybe related to Alzheimer's disease, however, too much zinc also can cause the beta amyloid deposit [2]. Some food, such as some legumes, mussels contaminated with demic acid, can stimulate the glutamate and is thought to be related to Alzheimer's disease [2]. Head injury, gender and educational level are also thought to be related to Alzheimer's disease [2]. Although some studies found higher amount of aluminum in Alzheimer's disease patient's brain tissue, it is not sure whether it has contribution to Alzheimer's disease or not [2].

**Research on diagnosis of Alzheimer's disease**

Since currently there is no way to treat Alzheimer's disease, early diagnosis plays a very important role. It gives time to control the symptoms, which can improve patients’ lives and decrease the work of care givers. Some researchers are focusing on the cognitive markers, which are the changes in mental abilities and personality [2]. Some examples are impaired visual memory, decreased consciousness and increased vulnerability to stress [2]. Others focus on biological markers and one studies showed that the tropicamide increase the pupil size of Alzheimer's disease patients more than normal [2]. Some studies use SPECT (single photon emission), MRI (magnetic resonance imaging) and
MRSI (magnetic resonance spectroscopy imaging) to scan the brain to try to find the cause of Alzheimer’s disease [2].

**Research on treatment of Alzheimer’s disease:**

There is also a lot of research on the treatment of Alzheimer’s disease. Acetylcholine decreases in Alzheimer’s disease patients and this leads to the possibility of cholinergic replacement therapy [2]. Drugs for this purpose are THA (tetrahydroaminoacridine), choline, lecithin, etc. [2]. However, the result was not good [2]. Some neurotrophic factors, such as NGF (nerve growth factor), BDNF (brain derived neurotrophic factor) and NT-3 (neurorophic-3) can regenerate the neurons and stimulate the growth of axons and dendrites, and thus were thought to be potential drugs for Alzheimer’s disease [1, 2]. Also, estrogen can interact with NGF and there are estrogen receptors in cholinergic neurons. There is a clinical trial underway to study an estrogen replacement technique [1, 2]. Since calcium plays an important role in Alzheimer’s disease, some researchers are looking for the calcium channel blockers [2]. Antioxidants such as vitamins C and E, beta-carotene, deprenyl and L-carnitine can eliminate the oxygen free radicals. There is an ongoing clinical trial in this area [1, 2]. Some studies use anti-inflammatory drugs to try to slow down the progress of this disease [2]. However, for some drugs, there is a very important problem, i.e., how to pass the blood-brain barrier, either through direct injection or using other substance to help them get in [2]. Many researchers have noticed that Alzheimer’s disease always has programmed cell death (apoptosis) and they are looking for ways to
inhibit this progress. Potential techniques for this purpose include using telomerase, reducing food intake, etc. [1]. There are also some researchers looking for an anti-amyloid Alzheimer's disease vaccine and they have gotten some positive results [1].

Managing symptoms is also very important in Alzheimer's disease research and this can improve the patient's ability to cope with daily living. This includes helping the patient keep independent and preventing disturbed behavior [2]. Some studies, such as improving cognitive function and improving functional abilities have promising results [2]. Drugs, such as anti-psychotics and antidepressants, as well as behavior management techniques, such as music and bright light, can help prevent disturbing behavior [2].

**Importance of Alzheimer's disease research and knowledge gap:**

Currently more than 4 million Americans have Alzheimer's disease. About 3% of people between 65 and 70 and about half of people with age 85 or older are suffering from Alzheimer's disease [1]. From age 65, the number of people having Alzheimer's disease doubles every 5 years [4]. The cost is $100 billion a year in USA [1]. It becomes more and more urgent to find the cause of Alzheimer's disease so that people can design drug to treat it and there are lots of researchers have dedicated to this. One problem in Alzheimer's disease research is that large amount of textual data such as research papers, books, clinical trial reports, etc., are generated. For example, using "Alzheimer's disease" as keyword to search MEDLINE, more than 33000 results can be found.
Furthermore, this number is increasing rapidly. To analyze these data and extract useful information from it is very time and labor consuming, even for an expert. A knowledge discovery tool, which can automatically extract information from these data, is urgently needed to help the researchers to obtain useful information in a short time and to improve the efficiency of Alzheimer’s disease research. However, currently this kind of tools does not exist. In this project, I developed a prototype system that can automatically detect the association of the terms related to Alzheimer’s disease from textual databases.
II. Background:

All the textual data are written in human language. Thus it is straightforward to use natural language processing (NLP) tools to analyze these data. NLP is to develop computer science method and software that can solve the problem related to natural language and great progress has been made recently. It needs the knowledge from linguistics, computer science, psychology, logic, and philosophy [6]. In bioinformatics, it has been used successfully to detect the relationship between genes and proteins [6, 7, 8].

Information extraction:

There are lots of textual data available, especially with the development of the Internet. Thus, the first step is to narrow down the data to the interested area. Currently there are several public databases available online that have a large amount of textual data. The first one is MEDLINE. To get related article from MEDLINE, we can use MeSH (Medical Subject Headings) terms, which can also be found from MEDLINE. There are also some specific databases that can be used. For example, the Combined Health Information Database (CHID) has specific topics such as Alzheimer's disease, AIDS, diabetes, cancer, etc. Some researchers also developed their own filter system. Ng and Wong designed a search engine named BioKleisli [9]. It can retrieve information from many online databases by providing a high-level query language [9].
Establish thesaurus:

A thesaurus must be established for the interested topic. The simplest way is using the thesaurus constructed by experts, for example, from public databases such as Genebank, PDB (protein database), and PubMed. However, the problem is, with the rapid increase in data, new terms, especially some genes and protein names are generated very quickly while the thesaurus is not updated accordingly. An annually updated thesaurus will miss much new information. Furthermore the latest information usually is the most important one to researchers. Some researchers, such as Ono et al. [8], constructed and maintained a keyword database by themselves. Obviously this is not an efficient way (although it is the most accurate) because it takes long time and needs expert knowledge. To address this problem, many tools are designed to automatically extract keywords from the textual data. Usually these tools fall into the following two categories: using lexical, syntactical and semantic analysis methods and using statistical methods [10].

One program that can identify the names and symbols of genes by using the lexical, syntactical and semantic analysis method was developed by Proux et al. [11]. This software first uses linguistic tools, which is doing tokenization, lexical lookup and disambiguation, to get the name of genes; then it does error recovery by using a dictionary and checking prefixes and suffixes; the last thing is to do a contextual lookup and validate remaining candidates [11]. 94.4% of terms can be identified and with the precision rate 91.4% by this program (the precision rate means that for all terms identified, how many are actually gene or
protein names) [11]. Using shallow parsing and noun phrase bracketing, Seikimizu et al. can find the keyword with a precision rate of 90% [6].

Other researchers found that there are some specific rules for the gene and protein names, for example, they are much different from normal English and usually have some specific forms; thus these names can be identified by using these rules [6, 9]. Hishiki et al. used character 4-grams to identify the gene name and it is very helpful combined with other tools [6]. Ng and Wong also set up a serial of rules, such as exclusion by standard dictionaries, inclusion with semantic clues, and inclusion with protein dictionaries to identify the name of proteins [9]. However, these rules have limited application to the names other than genes and proteins.

Andrade and Valencia used a statistical method to extract keywords from scientific text [10]. They used some proteins as examples and in their method. A set of background abstracts, which represents all abstracts that have information of all proteins, and a set of specific abstracts, which includes all abstracts that have information about one specific protein, were established; then the frequency of words in both set of abstracts were compared [10]. If the difference was significant, then the word was thought to be a key word [10]. They used a z-score to represent the correlation of the keywords to each protein and the higher the z-score, the higher the correlation [10]. Very good results were obtained in their three examples [10].

_Determine the association:_
Before determining the biological relation of two terms, first, it must be known whether they have relation or not. Stephen et al. used a statistical way to compute the relation of two terms [12]. They first converted every abstract into vectors of occurrence of every term; then the weighing factor of each term in each abstract was computed according to the term frequency, total number of term occurrence in all abstracts, and total number of abstracts; finally the association between two terms were calculated by the summation of the product of the two weighing factors of two terms [12]. Then user can define the threshold according to the expert knowledge or by using other methods [12]. Using some genes and their products as examples, they got precision rates from 61% to 100% according to the source of abstracts and references that they used as expert knowledge [12]. Most researchers just used the co-occurrence of two terms in one sentence to determine the association [6, 7, 8].

**Determine the meaning of association:**

The most important and difficult part is how to determine the biological meaning of the association between two terms. Most research focus on the relationship between genes and proteins. One reason for this may be that the relations between these terms are relatively simple. For example, the function of one protein to another protein can be simply being “activate” or “inhibit”. One method is presented by Ono et al. using surface clues based on particular patterns [8]. These patterns were created by using keyword, and the position of keyword, proteins names and other characteristic word constitute patterns, for
example, “protein A inhibits protein B” [8]. Then they used these patterns to search every sentence to see whether they have protein name and the specific patterns and determine the association [8]. Their method had a recall rate of 86.8% and a precision of 94.4% using their samples (the recall rate means for all association exist, how many can be identified) [8].

Instead of constructing the particular pattern, Ng and Wong established a database of function words such as “inhibit,” “suppress,” “regulate,” “induce,” “activate,” etc. [9]. They used these function words and protein names to search every sentence and determine their relations using pattern matching rules summarized by them, for example, “protein A” “function” “protein B”, “protein A “function of” “protein B”, “protein A” “which … Function” “protein B”, etc. [9]. Using “inhibit-activate” to search 26 articles about “effective human cyclin inhibitors”, they got 11 sentences matching their pattern rules and 16 protein-protein relations [9].

Sekimizu et al. developed a program that can automatically find the subject and object of frequently seen verbs such as “bind” and “inhibit” in medical papers, which were derived from MEDLINE [6]. They first searched these words in every sentence and if there was a verb, then determine their voice; then they looked left and right of the word to find the subject and object of the verb according to the voice [6]. This program can also deal with the cases such as “function” and “function of” [6]. However, in some complicated cases, the program gave false results because of parsing errors as indicated by authors. One example was if there were two “by” in a sentence that was in passive voice,
the program got confused and often gave the wrong results. Another example was if there was a clause and the frequently seen verb was in that clause, the object of sentence was thought as the object of the clause [6]. For all the verbs they had tested, the precision rate was from 67.8% to 83.3 %, with an average of about 73% [6].

Hishiki et al. used standard generalized markup language (SGML) tags to bracket the article [7]. Several tasks were designed to do different kind of works, for example, named entity (NE) task was to find the names and expressions; template element (TE) task can find the name of person and organization and determine the category and location of origination; scenario template (ST) task can be used to find the relationship for some specific terms; etc. [7]. Name entity tagger, which used decision trees, was used to find proper names and very good results were obtained by using only 100 training papers [7]. The system is still under development.

**Currently developed prototype system:**

Currently most of these natural language processing tools are used to find relations of genes and proteins. Several prototype systems have been developed to do such works and all of them can automatically detect the relation of genes and proteins. Ono et al.’s system maintained a keyword database, then used these keywords to establish particular patterns, then used these patterns to search against every article to get the relations of genes or proteins [8]. Ng and Wong’s system consists of three parts: BioKleisli, which can automatically search
the interested articles from many databases; BiolNLP, which uses function word pattern matching to obtain the association of proteins; BioJAKE, which is a graphic tool that can display the interaction pathway intuitively [9]. Sekimizu et al.'s method first selects the most frequent verb from MEDLINE abstracts which have 1 million words in total, then uses a shallow-parsing technique to parse every sentence to find the noun; the next step is using verbs and nouns to find the subject and object of the verb; if the noun is a protein name, then their relation can be determined [6]. All of these prototype systems have good results by using their testing samples. However, there are no such tools that can extract information from articles that is specific to some disease, such as Alzheimer’s disease.

**Intended research project:**

In my project, I developed a system that can automatically detect the association from terms related to Alzheimer's disease from MEDLINE abstracts. First I established a thesaurus about Alzheimer's disease. Due to limited time and the large number of terms that existed, I selected a suitable number of terms from the thesaurus that were already constructed by experts. Then I downloaded some abstracts that were related to Alzheimer's disease from MEDLINE. Using the method developed by Stephens et al. [12], I wrote a program that can convert each abstract to a vector that consists of weighing factors of each term in each abstract, and calculated the association of each term pair. The actual meaning of each term pair was determined by choosing the
sentences that include both terms. Also, graphs were drawn to represent the association of term pairs.

Finally a website was developed to make it very easy to check and understand the biological meaning of the association of each term pair. The final result was presented in this website. There were links to the definition of terms, the original abstracts (with all terms highlighted), and sentences that include both terms. Because large amount of term pairs were generated, to go through all of them took lots of time. Therefore this website was an interactive one, i.e., each time, only a small amount of term pairs that the user is interested in were presented and the user was asked whether they were interested in it or not. Each user's answer was recorded as their profile and next time the term pairs presented will be chosen according to their profile. Therefore, after limited times, only term pairs that the user is interested in will be presented and this can save user lots of time.
III. Methods:

*Overview of my system:*

The following graph is the overall structure of this system. It consists of four major parts. The first part is constructing the thesaurus (boxes 1 and 2); the second is processing the abstracts that are downloaded from MEDLINE (boxes 3 and 4); the third part is to calculate the association and find the annotation (boxes 5, 6, 7 and 8). The fourth part is to present the result in graph and constructing the interactive website (boxes 9, 10, 11 and 12).
Figure 1. Overall Structure.
Online resources of Alzheimer's disease:

Lots of articles about Alzheimer's disease can be found on the Internet. There are several major websites that have a large number of collections of Alzheimer's disease articles. One of them is Combined Health Information Database (CHID). This database has a specific topic about Alzheimer's disease, which has more than 6000 entries, including fact sheets, textbook chapters, journal articles, brochures, teaching manuals, directories, audiovisuals, bibliographies, program descriptions, monographs, newsletters, and reports from 1986 to present [13]. It is maintained by the Alzheimer's disease Education and Referral Center (ADECAR), which is a branch of National Institute on Aging. This database is updated every three months and is checked regularly to make the original articles correct and available [13]. This website also has an Alzheimer's disease thesaurus, which is also maintained by ADECAR. It has more than 1500 terms and for each term, there is a brief description about the category it belongs to and some related terms [14]. There is also a scope note for some terms [14].

Another major online resource is from PubMed, which is maintained by National Library of Medicine. It has more than 14000 articles about Alzheimer's disease from mid-1960 to today [15]. This website also maintains a vocabulary of biomedical terms called Medical Subject Headings (MeSH). MeSH is updated annually and like CHID thesaurus, it also has brief description for each term and the categories they belong to, which can be found using the MeSH Browser in their website [16].
The above two websites are free to every one. There is also another search engine for the articles of Alzheimer's disease named OVID [17]. It can be accessed in our university. The most important feature of this website is it can download full articles directly. For Alzheimer's disease, they have more than 6800 full articles [18]. All the PubMed articles can also be accessed from this website.

In my project, because of limited time, I did not use full articles. PubMed database includes many of CHID entries and has more than 14000 articles, thus I used abstracts from PubMed. Using "Alzheimer disease" to search MeSH major topic field for the articles with abstract, published in English, related to humans and published from 1986 to 08/2002, I obtained 14146 abstracts.

Establish the Alzheimer's disease thesaurus

Although automatically extracting keywords from the abstract can have the latest information, it also takes too much time to do; therefore I chose some terms from the Alzheimer's disease thesaurus in http://www.alzheimers.org/chid/dbframe.html [14]. Currently there are 1526 entries in this thesaurus. 353 of them are just abbreviations or synonyms of some other terms, so I removed them and got 1173 terms. 12 of them didn't belong to any categories and others belong to 22 categories, from which I chose the terms from the six categories in which I was interested. They were: chemicals and drugs, which had 114 terms; disease, symptoms and pathological processes, which had 169 terms; psychological and social behavior and characteristics,
which had 103 terms; therapeutics and patient care, which had 99 terms; endogenous chemicals, which had 31 terms; genetic elements, which had 9 terms. The total numbers of terms was 525.

**Text representation and getting the association**

The method used here was from Stephens et al. [12]. The first step was using all the terms to search against each abstract, thus this abstract can be represented by a vector of occurrence of each term; the number of abstracts that had each term was also counted; then a \(tf*idf\) algorithm was applied to get the weighing factor of each term in each abstract where \(tf\) is term occurrence and \(idf\) is inverse document frequency [12]. Thus the expression can be written as:

\[
W_{ik} = T_{ik} \times \log \left( \frac{N}{n_k} \right)
\]

Where \(W_{ik}\) is the weighing factor, \(T_{ik}\) is the occurrence of term \(T_k\), \(N\) is the total number of articles, \(n_k\) is the number of abstracts that have term \(T_k\) [12]. It can be seen from the above formula that the more abstracts in which a term appears, the lower the weighing factor because it is treated as the common term; on the other hand, the more the occurrence in an abstract, the higher the weighing factor [12].

The association strength of term \(k\) and term \(l\) is defined as [12]:

\[
\text{Association}_{kl} = \sum_{i=1}^{N} W_{ik} \times W_{il} \text{ where } k = 1 \ldots m, \ l = 1 \ldots m
\]
Obviously, if two terms appear in the same abstract, they have some kind of association. The more abstracts both of them appear in, the higher the association.

A Perl (Practical Extraction and Report Language, version 5.06) program was written to get association among terms. Perl is a very convenient and fast language for searching regular expressions; therefore, it is a perfect tool for calculating the term frequency and inverse document frequency. This program had two input files, one was all terms selected with one term in each line; the second was PubMed abstract, which was also one abstract in each line. This program used all terms to search against each line in PubMed abstract input file and changed each abstract to a vector of occurrence of each term. Then this program calculated the weighing factor and associations. All term pairs with association strengths were stored in an output file.

**Graphic representation of association**

A graph was drawn to represent the association of term pairs. This was done by using a trial version of SmartDraw, which was downloaded from http://www.smartdraw.com. The thickness of line between two terms represented their association strength. The thicker the line, the higher the association between these two term pair. Also terms that had strong relation were clustered together. This gave an intuitive way to view the association.
**Determine the biological meaning of association:**

The terms used in my project were from six categories and their relations were more complicated than the relation between genes and proteins. For example, the relation between symptoms and pathological processes, the therapeutic and patient care, etc. These relations need more verbs to describe them. Also, unlike many protein-protein interactions, which can be represented simply by using words such as “inhibit”, “activate”, etc., much information about drugs, such as “drug A have partly effect”, “drug B have strong effect”, are very important and need additional words other than verbs to describe them. Due to limited time, it was impossible to establish a database of these words or particular patterns of using these terms. Another important problem about denoting these terms was how to deal with a term pair that had totally different results in different articles, for example, some researchers thought aluminum was one of the cause of Alzheimer’s disease while others not. This is not usually seen in the interaction of protein-protein; furthermore, this kind of information cannot be described by just using a simple verb; therefore, other way must be used.

In my project, I used a program, which was also written by using Perl to try to find sentences that can typically represent the biological meaning of the term pair. This was realized by checking all the sentences that include the two terms from MEDLINE abstracts. The definition of a sentence was if it was ended with a period, then it was treated as a sentence. If there were many sentences for some particular term pairs, only one randomly selected one was used as the
denotations. The term pairs that didn’t have these kinds of sentences was marked as “not sure” and needed additional expert knowledge to denote them.

Establish the WebPages:

There was an interactive website to display all the results and their original sources. The website was constructed by using HTML4.0 and all CGI programs (Common Gateway Interface) were written by using Perl. This website was provided as a platform to facilitate users to check the result and get further information. The following were the functions of this website and the method about how I developed it.

First, all term pairs that had associations can be found in this website. And all their associations and annotations were displayed. There were two ways to check each term pair. The first one was through the main index page. All terms that have associations were displayed alphabetically here. Users can choose the term that they are interested in by directly clicking its name. The second way to see the association was through the interactive webpage. This webpage only presented users with term pair that they were interested in and I will talk it in detail later.

Second, this website also provided users all the original information. This was because of the fact that not all users were experts; furthermore, even for experts, there may be some new terms that they may not be familiar with or they want further information. Therefore, for each term pair, there was a link to all the definitions of these two terms that were from CHID thesaurus. Also, there were
links to every abstract that had these two terms with these two terms highlighted. If the term pair had annotations, there was a link to the source and other annotations (if any).

Third, since all the associations and annotations were generated automatically by the program, some expert’s knowledge should be needed to check the correctness. Therefore, for each term pair, there was a question about the correctness of it. Five choices were provided and they are:

1. Correct and straightforward, which means the association and annotation was correct and obvious. For example, the relation between placebo and any kind of drugs was very easy to see.
2. Correct, which means the association and annotation was correct but not obvious.
3. Correct but wrong annotation, which means there was association between the term pair, but the sentences the system chosen were not suitable annotation for the relation.
4. Not correct, which means there was no association between this term pair.
5. Not sure, which means the user didn’t know whether there should be relation between the term pair or not.

The user can only choose the best one.

Fourth, this website was an interactive one. As it was mentioned before, there were lots of term pairs with association. For some particular users, there may be not enough time to go through all of these terms; furthermore, users may
only interested in part of these term pairs, for example, a user may only want to see the associations between genetics elements and endogenous chemicals. It should be better if this system only presented some particular users with the term pairs that they are interested, i.e. this website should have “filtering” function. Thus, lots of time can be saved.

Finally, this website had the ability that let users to review all their answers and if they want, they can also edit their results.

In order to find what users are interested in and provides users only with the term pairs they are interested in as soon as possible, i.e. filtering terms, there was a profile for each user. This profile was adjusted according to the feedback of user every time this user visits this website. This is realized by keeping a record of profile for each user. Therefore, a username and password must be entered every time the user visits the website. However, in the first time when users visits the website, because there were no feedback from these users, this website will just give them a small amount of term pairs that are randomly selected. As I mentioned before, I have six categories, thus there were 21 kinds of category pairs and each term pair will belong to one of these 21 category pairs. So for the first time, the term pairs chosen will be according to the proportion of total number of term pairs that belong to each category pair in total number of term pairs. For example, if there are 5,000 term pairs in total and 500 of them belong to category pair genetic elements and endogenous chemicals, then this category pair has proportion of 10%; also suppose each time only 50
term pairs will be presented in this website, then only 5 term pairs will belong to genetics elements and endogenous chemicals.

For every term pair presented in the website, beside the question about the correctness, there was a question about whether the user was interested in this term pair or not. There were three choices for each question: very interested; interested; not interested. Because every term pair belongs to one of 21 category pairs, we can know which kind of category pair this user is interested in and this user’s answer will be recorded in this user’s profile. In the profile, if the user is very interested in a term pair, this category pair that this term pair belongs to will increase by 3. If the user is interested in a term pair, then the category pair will increase by 2. If not interested, then only increase by 1. Adding all category pair together we got the total number. Then the number of term pairs that will be chosen from each category pair is decided by its proportion in total number. Next time when this user visits this website, the term pairs will be selected according to this user’s profile. For example, suppose in someone’s profiles, they are very interested in 3 term pairs that belong to chemicals and drugs; interested in 4 term pairs that belong to genetic elements; not interested in 3 term pairs that belong to endogenous chemicals. Then the total number will be 3*3+4*2+1*3= 20. If 40 term pairs will be presented, then 3*3*40/20 = 18 will be chemicals and drugs; 4*2*40/20 = 16 will be genetic elements; 1*3*40/20 = 6 will be endogenous chemicals. The more feedback this websites gets from the user, the more accurate the user’s profile is. Therefore, after several times, only the term pair that the user is interested will be presented.
Check the correctness and test the website:

In order to check whether the association is correct or not, expert knowledge must be used. Also, it will be better to test the website's filtering function by several expert users. However, due to limited time and larger amount information I have, currently I cannot find such experts. Therefore, I went through all the term pairs with the help of original abstracts and definition of each term. Also I tested the website's filtering functions to see how accurate it can present the term pairs that the user is interested in.
IV. Results:

Associations among term pairs:

I had 525 terms and 137550 term pairs. I detected 3871 term pairs had association strengths larger than 0, which involved 298 terms. The highest association was 11128.03 and the lowest was 9.27. More than 92% term pairs had association strengths less than 500. About 62% were less than 100 and 40% were less than 50. The following is the distribution of associations.

Figure 2: Distribution of association
Graphic presentation of association:

Figure 3: Graphic presentation of association
Figure 3 is the graphic presentation of association for some term pairs that had association strengths about 1000. The thicker the line between each term pair, the higher the association strength.

**Correctness of association and annotation:**

I had gone through all 3871 term pairs with the help of original sources. The following was the result:

![Correct rate graph]

Figure 4: Correctness of association and annotation.
More than 90% associations were correct. About 60% term pairs also had correct annotation.

**Relations between association and correctness:**

Table 1 is average association strength for each category of answer. We can see, from correct and straightforward to not correct, the association strength was decreasing. This makes sense because for terms with higher association, there are more papers talking about them; therefore, there are higher possibilities to get the correct annotation.

<table>
<thead>
<tr>
<th>Category</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct and Straightforward</td>
<td>1690.65</td>
</tr>
<tr>
<td>Correct</td>
<td>203.56</td>
</tr>
<tr>
<td>Correct but wrong notation</td>
<td>129.77</td>
</tr>
<tr>
<td>Not correct</td>
<td>64.41</td>
</tr>
<tr>
<td>Not sure</td>
<td>102.49</td>
</tr>
</tbody>
</table>

Table 1: Relations between association and correctness
The design of the Website:

1. Log in.

In order to keep record of each user, user name and password is needed. Every user can obtain user name and password upon request.

![Log in page](image)

**Welcome! This is the website of associations among terms related to Alzheimer's Disease.**

If you have some comments, please email: diaux_teamirdo

Figure 5: Log in page.
2. Main page:

After the user logs in, the main page will be displayed. There is a brief introduction of the results and links to CHID and Pubmed. Also there are links to all association results. If the user wants to choose term by name, there is a link to main index. If the user chooses “see selected term pairs”, then the program will select term pair for the user.

![Image of Main Page](image_url)

We selected 525 terms from Alzheimer’s Disease Thesaurus, which is maintained by the Combined Health Information Database (CHID). These terms formed 137550 term-pair. We also used Mesh Major Topic "Alzheimer disease" and "has abstract" and "English" and "human" to search Pubmed and got 14890 abstracts. Then we used all term-pairs to search against these abstracts to obtain their association. We found that 3872 term-pairs, which include 298 terms, have association larger than 0. (Click here to see all the result, however, some Netscape users may have trouble to see it).

You can go to the main index to see the alphabetical list of these 525 terms and their associations.

Go to main index

We also select some term pairs. You can use the link below to see selected term pairs.

See selected term pairs

Figure 6: Main page.
3. Results table.

The following are the page that displays all term pairs and their association strengths.

![Figure 7: Association strengths table](http://www.cs.sju.edu/~bli-public/only2xv-MicrosoftInternetExplorer.png)
4. Display of association and annotation of each term pair:

Each term pair, their association strength and links to their original sources are in this page. The user can click each term pair to see the abstracts. Also if the annotation is other than "not sure", the user can also click it to see all the annotations. There are two kinds of questions: one is to check the correctness and the other is to see whether the user is interested in it or not.

Figure 8: Association and annotation of term pairs
6. Abstract page:

The following is a sample page that displays all abstracts that include the term pair and their definition. Each term are highlighted in the text.

Figure 10: Abstract page.
7. The annotation page:

The following is an example of annotation page.

Figure 11: Annotation page.
8. Edit results:

The following is a sample page for editing results.

The following are your results. If everything is O.K., please click confirm. If you want to make some change, please change by click the right answer then click confirm.

**ACTH : ACETYLCHOLINE**
- Correct and Straightforward
- Correct
- Correct but wrong notation
- Not correct
- Not sure
- Very interesting
- Interesting
- Not interesting

**AGITATION : ACETYLCHOLINE**
- Correct and Straightforward
- Correct
- Correct but wrong notation
- Not correct
- Not sure
- Very interesting
- Interesting
- Not interesting

**AGITATION : AGGRESSION**
- Correct and Straightforward
- Correct
- Correct but wrong notation
- Not correct
- Not sure
- Very interesting
- Interesting
- Not interesting

**ARTHRITIS : ANTIBODIES**
- Correct and Straightforward
- Correct
- Correct but wrong notation
- Not correct
- Not sure
- Very interesting
- Interesting
- Not interesting

**ARTHRITIS : ANTIGENS**
- Correct and Straightforward
- Correct
- Correct but wrong notation
- Not correct
- Not sure

Figure 12: Results editing page.
9. Main index page.

The following is the main index page. All terms are listed alphabetically.

Only terms that have associations are clickable.

Figure 13: Main index page
11. Detailed information about selected term pair.

For selected term pair, its association strength and links to its original sources are in this page. Again, there are two kinds of questions: one is to check the correctness and the other is to see whether the user is interested in it or not.

This is the association between MAPROTIENE and AMITRIPTYLIN.

MAPROTIENE: AMITRIPTYLIN, 8.66, It has been compared with amitriptyline, imipramine, clomipramine, desipramine, maprotiline, fluoxetine, bupropion, tranylcypromine, toloxatone, mianserin and amitriptyline in the treatment of depressive disorders.

Correct and straightforward Correct Correct but wrong notation Not correct Not sure

Very interested Interested Not interested

Submit Query...
Result of filtering:

The following is the result of filtering. It can be seen, at the beginning, because the profile is not accurate and there are some dips; however, after visiting the website about 15 times, almost all term pairs that are presented are what the user wanted.

Figure 16: Result of filtering
V. Conclusion:

This system is designed to automatically detect the associations among terms related to Alzheimer's disease and excellent results were obtained. More than 90% of associations were correct and about 60% of these term pair were given correct annotations. The correct rate of associations was very high. Although the correct rate of meaning of associations was kind of low, the algorithm was very simple and fast.

One of the best feature of this system is the extremely user friendly interface. This makes this system a very convenient and easy to use information extraction and knowledge discovery tool for studying Alzheimer's disease. The use of this website is very simple and there is a brief description in each page, therefore, no any additional training is needed. Because large amount of information exist, two ways about how to get useful information were provided. If users know the names of some particular terms, they can find them very easily through the alphabetical list of terms and all the terms have associations with them will be displayed. If the user wants to see some particular terms, the program will find them very quickly according to the user feedback and presents them to the user.

This system can be used as a great tool for knowledge discovery from textual data. Researchers in this area can find new information by using this system in a very short time and these kinds of information can be used to guide the further research. For example, a drug may or may not have some particular
kind of side effect. However, if this drug and this side effect have higher association, there is higher possibility that they may have some kind of relationship, thus, an experiment can be designed to test it according to this information. Furthermore, because the most important information about Alzheimer's disease are extracted and stored in this system, it can also be used as a convenient tool to improve the knowledge of Alzheimer's disease for beginners. For example, when I began to do this project, I just knew a little about Alzheimer's disease. However, during the time when I was checking the correctness of association and annotations, I obtained lots of knowledge about this disease, while it usually takes long time to get the same amount of knowledge from text book or original papers.
VI. Discussion:

Because of the limited time, the thesaurus I used was chosen from CHID. Although this thesaurus is updated very quickly and it is very accurate, it still cannot catch up with the latest development, while, on the other hand, with the rapid development of biotechnology, the latest information is often the most important one. Therefore, in the next step, a function unit that can automatically extract keyword from the latest publications will be developed. However, by using currently technology there are still lots of non-keywords included, or sometimes, false results are reported and we get wrong keywords, therefore, expert knowledge is still needed to check these new terms manually.

This system is a static system, i.e. it downloaded abstracts from MEDLINE and stored them in the in-house database, then processed them. Therefore, it only reflects the information that before some particular time. Again, this will miss the latest information. This problem can be solved by developing an interface or wrapper between this system and online resources, every time the online resources updated, the new information will be added into the system and process them together with former information. Thus, the latest information will be included.

I had 525 terms and there were 137550 term pairs. However, only 3817 term pairs had associations. One possible reason for this was that I just used the abstracts. Although this improved the processing speed, however, abstracts are just summary of the major topic of the paper, therefore, lots of information was
missed and this greatly limited the ability of knowledge discovery. Also, some papers, editorials, fact sheet, etc, do not have abstracts and are not included. In the future, a data warehouse about the full text of Alzheimer’s disease will be established and this system will process full text instead of just abstracts. Therefore, more complete information can be obtained.

I used the sentences that had both terms in them as the annotation. This was very simple and fast. However, for some term pairs, the sentences I got are very long and some sentences cannot reflect the relations between them. In order to find the annotation correctly and present them concisely, some natural language processing tools are needed to parse the sentences. Also, a thesaurus of key words that describe the relation should be established. Many researchers are trying to solve this kind of problem. However, there are still lots of problem.

For the information I got, some of them maybe outdated or not correct. For example, in early studies, aluminum was thought to be a cause of Alzheimer’s disease while recent studies show some negative results. However, because there are lots of papers about aluminum causing Alzheimer’s disease exist, I still got higher association strength and got the annotation that aluminum causes Alzheimer’s disease. One solution to this is maybe I should add some weighing factor to correct this. For example, the latest paper will have higher weighing factor and oldest one will have lower weighing factor, etc. This can make the result more accurate. However, lots of work needs to be done to determine the weighing factor.