# Discovering Fuzzy Association Rules from Patient's Daily Text Messages to Diagnose Melancholia

Yo-Ping Huang, Hong-Wen Chiu and Wei-Po Chuan
Department of Electrical Engineering
National Taipei University of Technology
Taipei, 10608 Taiwan
yphuang@ntut.edu.tw

Frode Eika Sandnes
Faculty of Engineering
Oslo University College
Oslo, Norway
frodes@hio.no

***Abstract*—With the constant stress from work load and daily life people may show symptoms of melancholia. However, most people are reluctant to describe it or may not know that they already have it. In this paper a novel system is proposed to discover clues from patient's interaction with psychologist or from self-recorded voice or text messages. A user friendly interface is provided for patients to input text messages or record a voice file by mobile phones or other input devices. A speech-to-text conversion software is used to convert voice mails to simple text files in advance. Based on the text files, a data mining model is used to discover frequent keywords mentioned in the text or speech files. The association rules can be used to help psychologists diagnose patients' degree of melancholia. Experimental results show that the proposed system can effectively discover melancholia keywords.**

***Keywords*—Data Mining, association rules, fuzzy model, word segmentation.**

## I. INTRODUCTION

Stress is a problem of modern society. People in different levels have faced different degree of stress. Stress may be reflected in the daily speech or in people's interaction with others. This makes some people prone to psychological symptoms such as melancholia and unconscious emotion reaction. Some people even do not know that they repeatedly mentioned some words in a short conversation with others. Consequently, a system that can help people early self-diagnose depression or help psychologist discover patients' unknown symptom is indispensible. In this paper data mining strategy is used to extract interesting association rules from patients' text messages.

Data mining is a technology developed to discover previously unknown but useful information from a large database. Two general criteria are applied to evaluate the interestingness of a discovered association rule, i.e., support and confidence. Support indicates the frequency of an item appearing in the transaction database. Confidence of X→Y implies the ratio of both itemsets X and Y simultaneous appearance to the frequency of X in the transactions. One commonly mentioned rule is "those who buy beers in the weekend will also buy diaper at the same time" [1, 2]. Without data mining technique, this important information may never be dug out from a supermarket database. The discovered association rules can be applied to promote a company's profit. For example, the supermarket can allocate both beer and diaper close to each other so that customers can easily find them.

This data mining concept is adopted in this study. Assuming a microphone is used to record a patient speech [3]. We then apply the data mining technique to find the associations among the spoken words. This is done by first converting the voice mails to text. Then, a term segmentation module is applied to segment words or phrases from the sentences. To perform the term segmentation, an on-line service module CKIP (Chinese Knowledge Information Processing Group) provided by Academia Sinica, Taiwan is used. Only nouns, verbs and adjectives in the segmented words are maintained as keywords and the rest are removed from the keyword list to simplify the analysis. After soliciting important keywords, they can be used to find the associations among them. For example, if both terms "daughter-in-law" and "unfilial" appear frequently in the text message, it may imply that the relationship between the mother and the daughter-in-law needs to be adjusted [4]. Besides, the emotional degree of a keyword can be used to diagnose the degree of melancholia [5]. For those with low depression or light symptom in psychology, the proposed system can be used to help patients detect and heal the symptoms earlier [6]. Of course, our work is heavily dependent on the determination of keywords' weights. In our study melancholia keywords are more valuable than other frequent words. Therefore, a depression degree of keywords measure is proposed.

This paper is organized as follows: methods for word segmentation, temporal data mining algorithm and fuzzy modeling are introduced in section II. The required hardware implemented in our experiments is illustrated in section III. Experimental results and analyses are presented in section IV. Conclusions and future research are given the final section.

## II. RELATED WORK

### *2.1 Word Segmentation*

Word segmentation is a crucial preprocessing step in natural language processing , such as voice recognition, document retrieval and machine translation. The results from word segmentation will affect the underlying discovery of frequent words in speech/text files. In English, a blank between two words indicate a natural boundary of separating two words. Compared to Chinese, English words are easier to segment.

978-1-4244-6587-3/10/$25.00 É2010 IEEE

Chinese sentences are composed of a series of words that make them hard to segment into meaningful phrases. Besides word segmentation it may be hard to assess the meaningfulness of a phrase. People's names, names of cities, terminologies and acronyms are examples that cannot be completely collected into the databases. These are the challenges for the word segmentation systems to be resolved. For example, in the Chinese sentence「我昨天晚餐吃漢堡(I ate hamburger for last night dinner)」we expect to segment the sentence into the meaningful term set: {我,昨天,晚餐,吃,漢堡}. Thus, a huge thesaurus database is needed to solicit meaningful words from sentences.

The CKIP word segmentation system [6] developed by Academia Sinica, Taiwan is exploited to extract meaningful words from sentences. The CKIP system provides the on-line word segmentation service that is based on XML data exchange mode. Users can use TCP Sockets to deliver their text messages to the segmentation system [7] to parse the marked-up XML results that include the segmented words and their corresponding term.

### *2.2 Fuzzy Set Theory*

Proposed by Lotfi A. Zadeh in 1965 [8] fuzzy set theory has extended its domain from theoretical study [9] to diverse areas such as auto-focus device in camcorders, water flow control in washing machines, temperature control in air-conditioners, auto-driving system in subway. Fuzzy set uses values between 0 and 1 to denote the degree of an element belonging to a set that is different from the traditional crisp set. We apply the fuzzy concept to denote the relative frequency of a segmented word to other words in a text file.

### *2.3 Association rules and FP-tree*

The idea of finding association rules from a large transaction database was first proposed by Agrawal et. al. [13]. Then, there are various algorithms presented to tackle the discovery of association rules from databases [10,11,14].

An association rule is expressed as: $X \rightarrow Y$, where X and Y represent two item sets. The rule indicates that if X were purchased then Y may be purchased at the same time. To discover the interesting association rules $X \rightarrow Y$, scanning the database to find occurring frequencies of both X and $X \cup Y$ is required. The appearance frequency of an item in the database is termed as support that means the frequency of the item appearing in the transactions. The support of $X \cup Y$ is the frequency of union sets X and Y appearing in the same transactions. If the support of $X \cup Y$ is divided by the support of X, the quotient is termed the confidence of $X \rightarrow Y$. If the confidence overpasses the preset threshold, then the association rule $X \rightarrow Y$ is treated interesting.

The first step to discover the association rules is to find out the large itemsets. There are two approaches to find the rules. One approach is to generate the candidate itemsets and the other does not generate candidate itemsets during the mining processes. Apriori is a representative that generates candidate itemsets. On the contrary, FP-tree did not rely on generating candidate itemsets to find association rules that result in a reduced execution time. The procedures to construct a FP-tree are described as follows:

(1) Scanning the database once and removing the duplicated items. Expressing those items with support greater than or equal to minimum support in itemset $L_1$.

(2) Ranking items in $L_1$ in descending order.

(3) Scanning the database again to construct a FP-tree and find all frequent itemset from the tree.

It is obvious that a database is scanned only twice in FP-tree to find frequent itemsets. Apriori algorithm, however, took much time in repeatedly scanning the database to find the same results. Fig. 1 is an example that illustrates how a FP-tree is constructed.

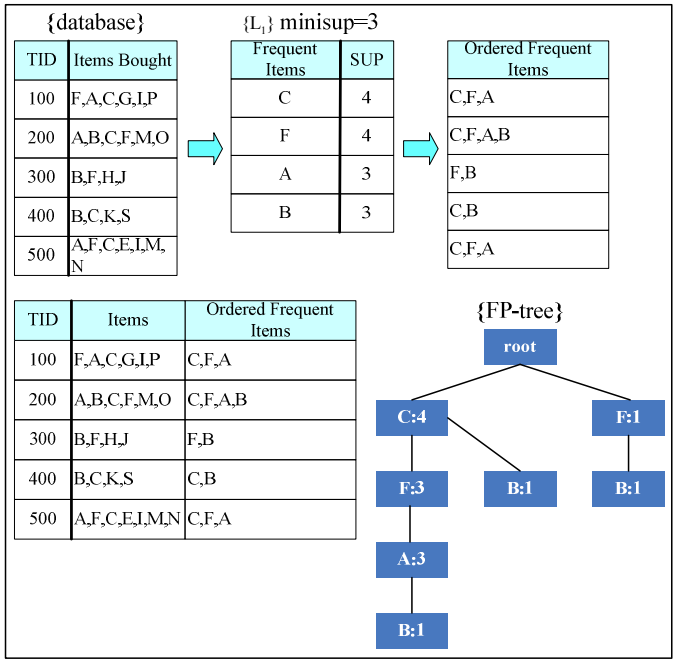


Fig. 1. A FP-tree example.

## III. SYSTEM FRAMWORK

To provide a simple way for users to upload their files to the system, either text or voice files are acceptable. A user can record a voice or text message either by computer or mobile phone. In the preliminary study, text messages are used for analysis. Then, the text messages are segmented to extract the keywords that show some degree of repetitions in speaking or texts or have some negative emotion. The analytical results can be used to remind the users themselves or their relatives to pay attention to the users speaking and emotion management.

The following describes the procedures to discover fuzzy association rules from a converted text message.

(1) Word segmentation: the CKIP word segmentation system developed by Academia Sinica, Taiwan is applied to extract keywords from text files.

(2) Keyword filtering: the extracted keywords are further filtered to remove undesired words and only nouns, verbs and adjectives are maintained.

(3) Weight calculation: the weight of a word is determined by its term frequency and its matching degree to the thesaurus of melancholia database. The weights of words are ranked in descending order to solicit the keywords in the text message.

(4) Fuzzy data mining: the keywords are analyzed by the fuzzy association rules [12,15] to find their associations and their fuzzy melancholia degrees.

The procedures to find fuzzy association rules among keywords are shown in Fig. 2.

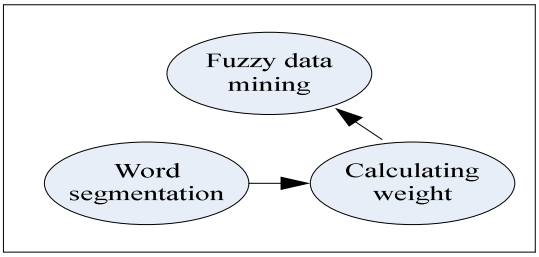


Fig. 2. Procedures to find fuzzy association rules among keywords.

The operational flow chart of the proposed system is shown in Fig. 3. For voice files they are converted to text messages before processing. A text message is then segmented into terms as shown in Fig. 4. Only nouns, verbs and adjectives in the segmented words are maintained as keywords and the rest are removed from the keyword list. The purpose here is to remove the platitude or uninteresting words and make the analytical results more accurate.

After segmenting words from sentences, the next step involves soliciting representative keywords. First, the term frequencies of keywords must be ranked in descending order. However, the term frequency is not the only criterion to determine its importance in the text message. For example, if both "daughter-in-law" and "unfilial" were frequently mentioned in the text, then there is no doubt that both are treated as important keywords. In reality, low term frequency words may also be critical keywords. For example, "suicide" may appear once in the whole text message but it may play the decisive role in determining that the patient has melancholia. A new strategy is proposed to calculate the weight for a keyword.

Before determining the weight of a keyword, a melancholia thesaurus is given in Table 1. In the thesaurus, each keyword is given a figure to denote its degree of emotional depression. The following is explanation for this thesaurus. There are many reasons to cause symptoms of depression, for example, losing loved ones, losing job and getting ill. Changing in thinking is a common symptom such as patients feel that there is no hope in life and have no value in society. According to above possible symptom, our research defines several keywords of depression to put the keywords in depression thesaurus such as "sad" and "disappointment". There are more serious symptoms that will have very negative thoughts, for example, often thinking of death and attempting to commit suicide. These patients think death can solve everything. This level of thinking is often very severe symptoms of depression. So our research defines other keywords of depression such as "suicide" and "dead". Then the weight of "suicide" and "dead" must higher than "sad" and "disappointment" because of that symptom is more serious. Of course, the depression degree should be derived from the domain experts. In this study, only simulation figures are given for experiments.

Assuming the term frequency and the depression degree in the thesaurus each has 50% of importance to the weight. Take the term of "daughter-in-law" for example. Its weight is calculated as $0.135\times0.5+0\times0.5=0.068$ due to the fact that "daughter-in-law" did not appear in the melancholia thesaurus and its depression degree is zero. For the term of "suicide", it weight is $0.027\times0.5+0.07\times0.5=0.049$ due to its depression degree in the thesaurus of 7%. By introducing the depression degree, a seldom appearance term may take a better position in the ranking list of importance as shown in Table 2.

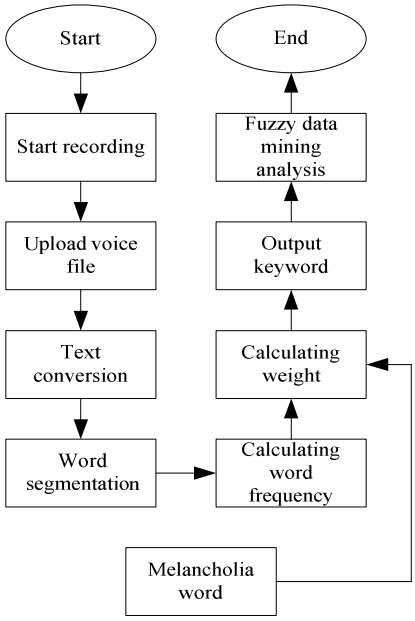


Fig. 3. Operational flow chart of the proposed system.

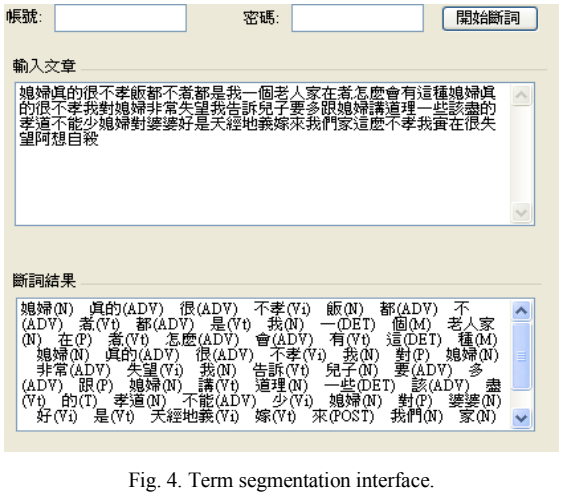


Fig. 4. Term segmentation interface.

Table 1. A melancholia thesaurus.

| Melancholia term | Depression degree |
|---|---|
| 自殺(Suicide) | 7% |
| 死(Dead) | 6% |
| 悲哀(Sad) | 2% |
| ⋮ | ⋮ |
| 失望(disappointment) | 1% |

Table 2. Term frequency and its calculated weight.

| Term | Term freq. | Percentage | Weight |
|---|---|---|---|
| 媳婦 (Daughter-in-law) | 5 | 0.135 | 0.068 |
| 我(I) | 4 | 0.108 | 0.055 |
| 不孝 (Unfilial) | 3 | 0.081 | 0.041 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 自殺 (Suicide) | 1 | 0.027 | 0.049 |

## IV. Experimental Results and Analysis

A melancholia database is used to record patients' mood or emotion. In the following experiments only text messages are solicited for analysis and the other information is discarded to protect patients' privacy. The following five negative emotion messages were solicited for the experiments. One sample text message is "My daughter-in-law is very unfilial and does not cook. I, an elder, cook myself. How come there is such a daughter-in-law who is very unfilial. I am very disappointed about her. I told my son that he should say something to her and be filial. It is the golden rule that a daughter-in-law should be filial to the mother-in-law. After marrying to my son she is very unfilial and I feel so disappointed. I want to commit suicide." as shown in the upper part of Fig. 4. Note that the original Chinese text is: "媳婦真的很不孝飯都不煮都是我一個老人家在煮怎麼會有這種媳婦真的很不孝我對媳婦非常失望我告訴兒子要多跟媳婦講道理一些該盡的孝道不能少媳婦對婆婆好是天經地義嫁來我們家這麼不孝我實在很失望阿想自殺."

After combining both term frequency and melancholia degree into calculation, the weights are given in Table 2. In our experiments, the weight for the parent node is also considered. A parent node is considered to emphasize the role of keywords in melancholia database playing in the analysis. Thus, weight for the parent node T1 in Table 3 is the addition of weights attached to the corresponding keywords in the melancholia database. For example, (suicide, 0.049)+(disappointment, 0.032)=(T1, 0.081) for the first message in Table 3.

Next, we will define membership functions for the weights of solicited keywords as shown in Fig. 5. In our experiments, three labels, Low, Medium and High, are defined for the variable "weight." This step fuzzifies the weights into the mapping linguistic variables that are easier to be verified by psychologist. The fuzzy sets converted from membership functions are given in Table 4.

Since each weight may be mapped to more than one linguistic term the one with higher membership degree is chosen as the linguistic term for the keyword. For example, the keyword "unfilial" has a calculated weight of 0.041 in the first message. Then, the weight of 0.041 is transformed to Medium because it has membership degrees of 0.45 and 0.55 in Low and Medium, respectively. As a result, each solicited keyword is accompanied a linguistic term to denote its depression degree. Based on the fuzzy degrees given in Table 4 we can sum up the depression degree for each keyword and its accompanied linguistic term for a patient in a certain period of time. For example, "unfilial and medium" has a total depression degree of 0.8 summing from the first and the 5th messages. If the threshold of depression membership degree is set to 0.8, then some unqualified keywords can be removed from succeeding analysis as shown in Table 5. The candidate keywords are then used to construct a FP-tree as shown in Fig. 6.

Table 3. Weights for the solicited keywords.

| RD | Solicited Keywords |
|---|---|
| 1 | (我, 0.055) (媳婦, 0.08) (不孝, 0.041) (自殺, 0.049) (失望, 0.032) (T1, 0.081) |
| 2 | (媳婦, 0.045) (悲哀, 0.042) (離婚, 0.037) (T1, 0.042) |
| 3 | (媳婦, 0.06) ( 自殺, 0.055) (離婚, 0.04) (悲哀, 0.031) (死, 0.055) (T1, 0.141) |
| 4 | (命苦, 0.04) ( 失望, 0.033) (離婚, 0.04) (自殺, 0.032) (死, 0.045) (T1, 0.11) |
| 5 | (死, 0.042) ( 不孝, 0.065) ( 媳婦, 0.06) (T1, 0.042) |

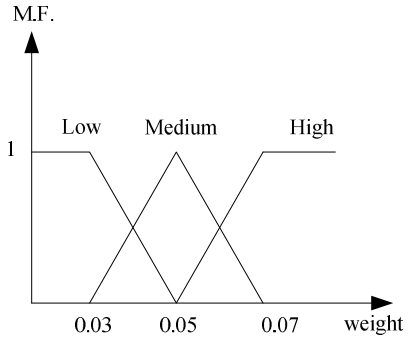


Fig. 5. Membership functions for weight.

Table 4. Fuzzy sets converted from membership functions.

| RD | 我 | | | 媳婦 | | | 不孝 | | |
|---|---|---|---|---|---|---|---|---|---|
| | L | M | H | L | M | H | L | M | H |
| 1 | | 0.75 | 0.25 | | 0.10 | 0.90 | 0.45 | 0.55 | |
| 2 | | | | 0.25 | 0.75 | | | | |
| 3 | | | | | 0.50 | 0.50 | | | |
| 4 | | | | | | | | | |
| 5 | | | | | 0.50 | 0.50 | | 0.25 | 0.75 |
| counts | 0.00 | 0.75 | 0.25 | 0.25 | 1.85 | 1.90 | 045 | 0.80 | 0.75 |

| RD | 自殺 | | | 失望 | | | 死 | | |
|---|---|---|---|---|---|---|---|---|---|
| | L | M | H | L | M | H | L | M | H |
| 1 | 0.05 | 0.95 | | 0.9 | 0.10 | | | | |
| 2 | | | | | | | | | |
| 3 | | 0.75 | 0.25 | | | | | 0.75 | 0.25 |
| 4 | 0.90 | 0.10 | | 0.85 | 0.15 | | 0.25 | 0.75 | |
| 5 | | | | | | | | | |
| counts | 0.95 | 1.80 | 0.25 | 1.75 | 0.25 | 0.00 | 0.25 | 1.5 | 0.25 |

| RD | 離婚 | | | 悲哀 | | | 命苦 | | |
|---|---|---|---|---|---|---|---|---|---|
| | L | M | H | L | M | H | L | M | H |
| 1 | | | | | | | | | |
| 2 | 0.65 | 0.35 | | 0.45 | 0.55 | | | | |
| 3 | 0.5 | 0.50 | | 0.90 | 0.10 | | | | |
| 4 | 0.5 | 0.50 | | | | | 0.50 | 0.50 | |
| 5 | | | | | | | | | |
| counts | 1.65 | 1.35 | 0.00 | 1.35 | 0.65 | 0.00 | 0.50 | 0.50 | 0.00 |

| RD | T1 | | |
|---|---|---|---|
| | L | M | H |
| 1 | | | 1.00 |
| 2 | 0.45 | 0.55 | |
| 3 | | | 1.00 |
| 4 | | | 1.00 |
| 5 | 0.45 | 0.55 | |
| counts | 0.90 | 1.10 | 3.00 |

Table 5. Depression membership degrees for solicited keywords.

| Keyword | Count |
|---|---|
| 我.M | 0.75 |
| 媳婦.H | 1.9 |
| 不孝.M | 0.8 |
| 自殺.M | 1.8 |
| 失望.L | 1.75 |
| 死.M | 1.5 |
| 離婚.L | 1.65 |
| 悲哀.L | 1.35 |
| 命苦.L | 0.5 |
| T1.H | 3 |

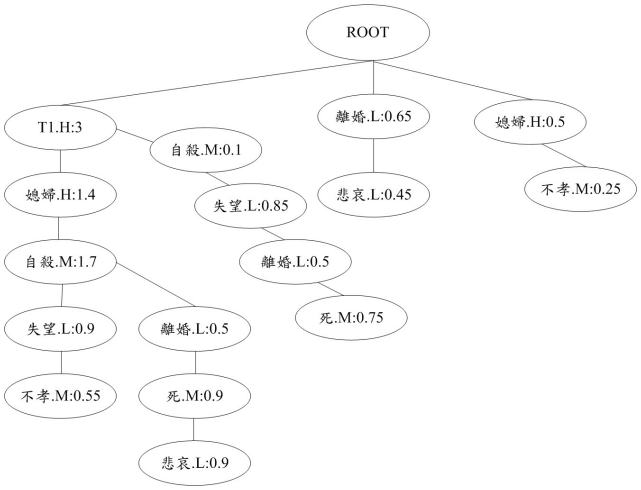


Fig. 6. FP-tree for the solicited keywords.

From the constructed FP-tree we can find the association rules among the keywords. For example, if we want to find the association rule (daughter-in-law,H)→(T1,H), then its confidence value can be calculated as follows:

$$\frac{\sum((daughter-in-law,H)\cap(T1,H))}{\sum(daughter-in-law,H)}=\frac{1.4}{1.9}=73\%\cdot$$

The figure given in the numerator of the above equation is derived from taking the minimum operation on the associated terms as shown in Table 6. The derived association rule, (daughter-in-law,H)→(T1,H), implies that if the term "daughter-in-law" belongs to the linguistic term "High" then it is probable that the patient shows some degree of depression. This is because (T1,H) indicates the high degree of depression in the melancholia thesaurus. Besides, based on this association rule, we can guess that the mother and her daughter-in-law have some problems to get along with in the daily life. Consequently, the mother often mentioned the term "daughter-in-law" in her messages. Also, another association rule, (daughter-in-law,H)→(suicide,M) with a confidence of 73%, is found from our analysis. The analytical results can help detect a patient depression degree and take any necessary action to protect from commit suicide. Other association rules discovered from the given example are listed in Table 7.

Table 6. The minimum operation on the associated terms (daughter-in-law, H) and (T1, H).

| RD | 媳婦.H | T1.H | 媳婦.H∩ T1.H |
|---|---|---|---|
| 1 | 0.9 | 1 | 0.9 |
| 2 | 0 | 0 | 0 |
| 3 | 0.5 | 1 | 0.5 |
| 4 | 0 | 1 | 0 |
| 5 | 0.5 | 0 | 0 |
| counts | | | 1.4 |

Table 7. Association rules from the given example.

| |
|---|
| 媳婦.H → T1.H=1.4/1.9=73% |
| 媳婦.H → 不孝.M=0.8/1.9=42% |
| 媳婦.H → 自殺.M=1.4/1.9=73% |
| T1.H → 離婚.L=1.55/3=51% |
| T1.H → 死.M=1.5/3=50% |
| T1.H → 自殺.M=1.8/3=60% |

## V. Conclusion and Future Work

In this paper FP-tree is used to discover association rules from patient's text messages. An user interface is devised for patients to upload their text messages or voice mails. For those voice mails, a speech-to-text module is expolited to make the conversion. A CKIP module provided by Academia Sinica, Taiwan is used to segment words from sentences and remove trivial words from succeeding analysis. A melancholia thesaurus is also established to take the emotional or depression keywords into consideration. Both term frequency and melancholia keywords are considered to calculate a term's weight. Fuzzy membership functions are also defined to fuzzify a term weight. A keyword is attached with its corresponding linguistic term to denote its degree of depression. Based on the paired keywords and linguistic terms, FP-tree is used to discover the association rules among them. The discovered results can help psychologist diagnose their patients' degrees of melancholia and take any necessary action in advance to prevent from commit suicide. Besides, the analytical results can help patients themselves adjust their social relationships with friends or relatives, or remind themselves paying more attention to their psychological conditions.

The current results are based on the analysis from text messages. Using speech-to-text module to convert voice mails into text messages or analyzing the voice mails directly is a great challenge in the continuing study. How to segment terms from sentences and then how to determine meaningful keywords or phrases from terms are also deserved further study in the future.


## Acknowledgment

This work is supported by National Science Council, Taiwan under Grants NSC97-2221-E-027-034-MY3 and joint project, No. 211-1-2, between National Taipei University of Technology and Mackay Memorial Hospital. The on-line service of term segmentation system provided by Academia Sinica, Taiwan is also highly appreciated.


## References


**[1]** J.-S. Yoo, "Similarity-profiled temporal association mining," *IEEE Trans. on Knowledge and Data Engineering*, vol. 21, no. 8, pp.1147-1161, August 2009.

**[2]** D. Perera, J. Kay, I.W. Koprinska and K. Yacef, "Clustering and sequential pattern mining of online collaborative learning data," *IEEE Trans. on Knowledge and Data Engineering*, vol. 21, no. 6, pp.759-772, June 2009.

**[3]** L. Buera, A. Miguel, O. Saz and A. Ortega, "Unsupervised data-driven feature vector normalization with acoustic modal adaptation for robust speech recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 2, pp.296-309, Feb. 2010.

**[4]** D.-A. Chiang and C.-T. Wang, "The cyclic modal analysis on sequential patterns," *IEEE Trans. on Knowledge and Data Engineering*, vol. 21, no. 11, pp.1617-1628, Nov. 2009.

**[5]** J. Li, A.W.C. Fu and P. Fahey, "Efficient discovery of risk patterns in medical data," *Artificial Intelligence in Medicine,* vol. 45, no. 1, pp.77-89, Jan. 2009.

**[6]** Chinese thesaurus on-line service, http://ckipsvr.iis.sinica.edu.tw/

**[7]** renjin's blog, http://renjin.blogspot.com/2009/04/ckip-client-for-net.html#comment-form

**[8]** L.A. Zadeh, Fuzzy sets, *Information and Control*, vol. 8, no. 3, pp.338-353, 1965.

**[9]** Z. Huang, T.D. Gedeon and M. Nikraves, "Pattern trees induction: a new machine learning method," *IEEE Transactions on Fuzzy Systems,* vol. 16, no. 4, pp.958-970, August 2008.

**[10]** X. Shang, K.U. Sattler and I. Geist, "SQL based frequent pattern mining without candidate generation," in *Proc. of the 2004 ACM symposium Conf. on Applied computing,* Nicosia, Cyprus, pp.618-619, 2004.

**[11]** C.-M. Cha and Y.-C. Tai, "An SQL-based improvement of the FP-tree construction technique," *Information Management Research*, vol. 6, pp.31-46, July 2006.

**[12]** T.-P. Hong, K.-Y. Lin and S.L. Wang, "Fuzzy data mining for interesting generalized association rules," in *Proc. of the 3rd Int. Conf. on Machine Learning and Cybernetics*, Shanghai, China, pp.26-29, August 2004.

**[13]** R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large database," in *Proc. of ACM SIGMOD Int. Conf. on Management of Data*, Washington D.C., U.S.A., pp.207-216, May 1993.

**[14]** Y.-P. Huang, L.-J. Kao and F.E. Sandnes, "Efficient mining of salinity and temperature association rules from ARGO data," *Expert Systems with Applications*, vol. 35, no. 1-2, pp.59-68, Aug. 2008.

**[15]** Y.-P. Huang, L.-J. Kao and F.E. Sandnes, "Predicting ocean salinity and temperature variations using data mining and fuzzy inference," *Int. Journal of Fuzzy Systems*, vol. 9, no. 3, pp.143-151, Sept. 2007.