AN ARTICLE REORGANIZATION MODEL BASED ON EMOTION IMPLIED IN FORUM ARTICLES

***Shih-Ting Yang*[1*] *and Yao-Chang Tsai*[2]**

[1]Department of Industrial Engineering and Systems Management, Feng Chia University
40724 Taichung, Taiwan (R.O.C)

[2]Department of Information Management, Nanhua University
62249 Chiayi, Taiwan (R.O.C)

Email: styang@fcu.edu.tw[1](corresponding author), 10169003@nhu.edu.tw[2]

DOI: https://doi.org/10.22452/mjcs.vol33no3.4

ABSTRACT

At present, the virtual forums are free and convenient speaking platforms, according to the speaking specifications and through the examination of forum administrators, the forum users can publish various articles easily. However, as the number of users increases, it is difficult for the forum administrators to check them one by one, and return all the articles to be revised. Also, the publishers (article providers) may write violative words unconsciously, so the article writers infringe the forum specifications with the publication, and shall revise the violative articles. Therefore, this paper develops an Article Reorganization Model based on Emotion Implied in Forum Articles including Article Expressed Emotion Determination Module and Publisher's Article Statement Reorganization Module. The first module deduces the emotion type of articles by analyzing representative events of articles, creating emotion word membership coefficient, analyzing similar statements and analyzing emotion probability and stable value. The second module uses integrated semantic similarity analysis, review score analysis and multiple combined sentences establishment, the article's statement structure can be reorganized. This paper builds a Web-based system and a real-world case is applied to confirm the feasibility of this methodology. For forum administrators, violative articles can be extracted rapidly from the articles with specific emotions and the violative statements will be reorganized. For article writers, the statements of violative articles can be revised automatically to save the time for revising violative articles.

Keywords: Virtual Forum, Sentences Similarity, Emotion Determination, Article Statement Reorganization

1.0 INTRODUCTION

Before the forum was popular, most of users discussed only with related acquaintances about their opinions on things. However, the free and convenient speaking platform of the present forums has become a main channel for people to declare their views and opinions ([3], [7]), such as "Mobile01", "Gamer", and "Eyny Forum", various articles can be published easily after they are approved by speaking norms and forum administrators [1]. In fact, as the number of forum users increases gradually, the article lengths and contents published by article providers are mostly different. Thus, the forum administrator must review all articles according to the speaking specifications and the word usage to filter violative articles to maintain the article quality of forum [10]. Most of forums have administrators for inspecting violative articles, but a lot of articles are put in the forums, the forum administrators are hard to examine all articles one by one [15]. In addition, the article providers with special opinions and views on specific things may write extreme or critical words in the articles unconsciously due to personal carelessness, violating the rules, so that the articles are removed by the forum platform or administrators.

According to the aforesaid content, the forum administrator only removes all the violative articles following the specifications of forum, the article review mechanism of forum even can use the existing Text-Matching mechanism to remove the articles with violative words directly [18]. The administrator can read all violative articles one by one in his own view, to obtain articles with particular views, and then the administrator makes revision or asks the article provider to make revision. However, this action costs much manpower and time, so that the efficiency of execution is poor. To sum up, the existing operation model (AS-IS Model) can be shown in Fig. 1.1. The motivations and purposes of this paper can be reduced to the following two points:

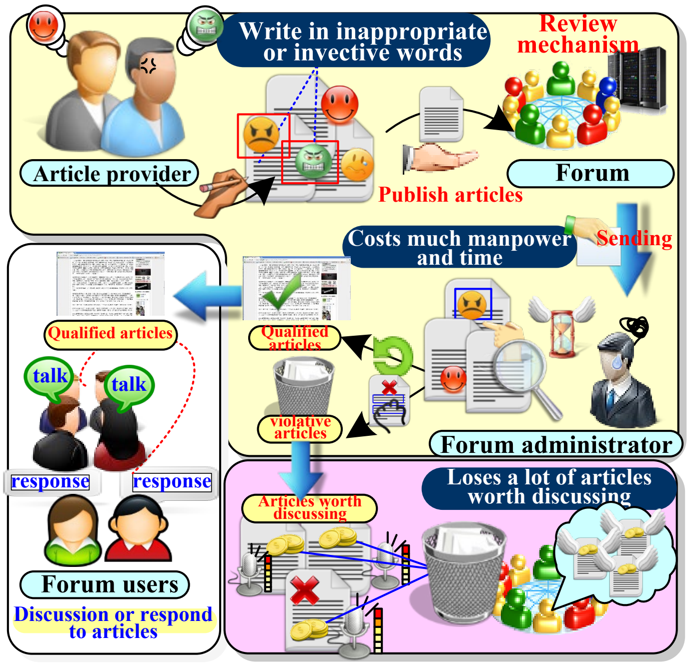


Fig. 1.1: The existing operation model (AS-IS Model)

1. When an article provider wants to express his thought, he may write in inappropriate or invective words unconsciously due to personal carelessness, violating the speaking specifications of forum.

2. It is difficult for the forum administrator to revise inappropriate words in the article content one by one. The administrator only removes violative articles following the specifications of forum, so that the forum loses a lot of articles worth discussing.

In order to assist the forum administrators to obtain the articles with discussion values from violative articles to maintain meaningful articles, this paper conducts an Article Reorganization model based on Forum Article Implied Emotion, analyzes the articles with extreme words, and reorganizes the statement content for this type of articles, so as to avoid being removed as violating the speaking specifications of forum. The expected operation model (TO-BE model) of this paper can be shown in Fig. 1.2, the key points of this paper are reduced to the following two points:

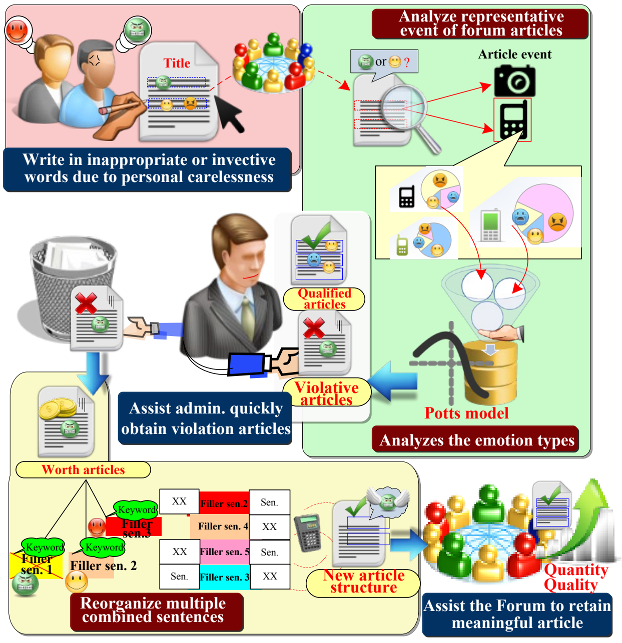

Title
Write in inappropriate or invective words due to personal carelessness
Analyze representative event of forum articles
or ?
Article event
Qualified articles
Violative articles
Potts model
Analyzes the emotion types
Assist admin. quickly obtain violation articles
Worth articles
Keyword
Keyword
Keyword
Filler sen. 1
Filler sen. 2
Filler sen.3
XX
Filler sen.2
Sen.
Filler sen. 4
XX
XX
Filler sen. 5
Sen.
Sen.
Filler sen. 3
XX
New article structure
Quantity
Quality
Reorganize multiple combined sentences
Assist the Forum to retain meaningful article

Fig. 1.2: The expected operation model (TO-BE model)

1. Analyze emotion type of article

When an article provider wants to express his opinion on specific things, he often uses different emotional words to enhance the mood of statements. However, when the article provider uses extreme emotional words, which are likely to offend the specifications of forum. Therefore, in order to obtain the article provider's emotion type, this paper uses the representative statements of article and analyzes the similarity between statements, and uses emotion analysis technique to analyze all of similar statements to obtain the emotion type of article.

2. Reorganize article with extreme emotion and words and statements

When the article provider has malicious or extreme emotion, this paper analyzes the grade score of content for this type of articles, the article with low grade score means the smoothness and connectivity of the content are bad. Therefore, this paper analyzes the important statements in the article and uses multiple combination statements to reorganize the content of article. Then the extreme wording in the article violating the forum specifications can be avoided to assist the forum administrator to keep meaningful and non-violative articles.

tIn order to assist the forum administrator to keep the articles worth discussing and not violating forum specifications, this paper analyzes the emotion types of articles to calculate the probability of articles with malicious or extreme emotion, and analyzes the review score of article content to know the statement coherence. Finally, this paper uses the combined statements of important statements to reorganize the content of articles. The purpose of this paper is to assist the forum administrators to obtain violative articles rapidly, and to reorganize the content of the violative articles but with particular opinions (enhancing articles discussed by forum users continuously), so as to keep the meaningful articles.

2.0 LITERATURE REVIEW

The core research subjects involved in this paper include "discussion about virtual forum management system", "forum article analysis" and "discussion about article writer behavior". Related studies are reviewed below.

2.1 Discussion about virtual forum management system

Wang et al. [5] used the same knowledge sharers and similar contextual messages to collect similar knowledge in the same knowledge base, classified the knowledge base into different formats and functional levels. The study classified all knowledge sharers in the knowledge base of the same level and format as one class, so as to use real-time communication tools to assist the knowledge extractors in communicating with sharers directly. Fang et al. [16] surveyed 143 IT forum members by questionnaire, and analyzed the fairness of forum feedback members (e.g. excitation mechanism), how the members shared knowledge and member-to-member, member-to-forum trust and knowledge reliability in the questionnaire. The feedback mechanism of forum and the trust of forum members in the forum can be obtained from the analysis result, which are the factors influencing the members' continuous and participating in knowledge sharing.

At present, the virtual communities have excessive knowledge gradually, so it is difficult for the forum administrators to supervise and classify knowledge. Therefore, Gu and Grossman [15] constructed a knowledge sharing platform for the LAN with higher knowledge exchange volume. The platform uses P2P technique to enable the knowledge extractors and sharers to share knowledge at any time, and uses distributed knowledge analysis method to classify the knowledge. The administrator can create another knowledge sharing platform for the domains of high knowledge quantity, so that the knowledge extractors and sharers of the same domain can share more centrally. Therefore, the platform can assist the administrators to integrate and filter the exclusive domains of knowledge from the massive knowledge, and facilitate the users to extract and share the required knowledge. In addition, in order to further increase the efficiency and effectiveness of knowledge management, Lai [9] proposed a theoretical model of knowledge engineering based on knowledge management (Knowledge Management through Knowledge Engineering; KMKE). The model uses four hierarchical mechanisms, which are knowledge modeling, knowledge verification mechanism, knowledge stratification storage and knowledge update mechanism, for knowledge management. By the operation of four hierarchical mechanisms, different types of knowledge can be presented by similar expressions, and the verification mechanism verifies the knowledge presented in formalized semantics. All knowledge can be converted into similar types and classified and stored the knowledge in the system by layers to assist the knowledge managers to manage the massive and frequently changed knowledge. Moreover, in different language system development, Rehmam et al. [27] constructed the idea of interface designing for ASL (Arabic Scripted Language). Firstly, this study presents a novel idea of designing an ASL interface for desktop application e.g., databases, applied for Urdu language. Secondly, American Standard Code for Information Interchange (ASCII) codes are used for mapping the keystrokes to the Urdu character's images using Phonetic keyboard styles. Then, the Urdu characters are contained in these fonts is mapped through ASCII codes. Therefore, this study developed an algorithm for Urdu desktop controls. This study can provide the concept to new researchers and software developers to develop interface not only for Urdu but also for any ASL. Furthermore, in speech recognition area, Rehmam et al. [26] proposed a system for speaker independent speech recognition of isolated words from the oriental languages. The study combined both the DWT (Discrete Wavelet Transform) with FFANN (Feed-forward Artificial Neural Network) to recognize isolated words. The system performance presented high accuracy for two and five classes. Therefore, the proposed study and system can be utilized as a communication interface to computing and mobile devices for low literacy regions.

2.2 Forum article analysis

In terms of forum article analysis, Ko and Seo [19] proposed an abstract formation model of composite statistical continuous virtual statements. Firstly, two continuous statements of document were combined into a virtual statement by using Bi-gram technology, and the time weight between two statements was added to the user slide window experimentally. Secondly, the keywords were screened by TF-IDF algorithm and the keywords were used as criteria of statement similarity measurement. Finally, the most important and the most representative statements can be obtained to form the article abstract. The representative statements can be obtained from multiple articles and unstructured articles effectively to form the abstract by using the method. In addition, Hao et al. [24] proposed a problem-based corpus creation method. The study used semantic knowledge base to tag all statement target words in the training data, the semantic frame, the statement type and syntax. The study defined eight problem types, and defined the problem types of all statements. After that, the semantic frame can be created according to the repeatability of target words and multiple semantic relations, so that the definitions of target words and statements were divided more clearly. Finally, the statement with multiple target words was labeled, the statement was linked to other target words. The words can be linked to statements, ambiguous words and problem types by using the aforesaid method, so as to form a complete corpus to assist the users to give the demanders appropriate response based on the corpus.

Ge and Chen [13] firstly classified the feature words, such as numerals and terms, of compositions and filtered the stop words to restore the roots of important words, and then used the feature words of all compositions and the roots of feature words to calculate the vector of features by Term Frequency-Inverse Document Frequency (TF-IDF) to obtain the feature weights of all features. Then, the cosine function of Vector Space model was used to analyze the similarity among all compositions. Finally, the study used hierarchical agglomerative clustering method to cluster similar compositions as one class, and recalculated the similarity among all compositions in the group. Therefore, the compositions can be clustered in the same cluster till one cluster was left to form a clustering tree; then, the composition content not classified as any cluster was regarded as an excursive composition. On the other hand, Liao et al. [2] proposed a statement ranking method to assist the users to create an appropriate statement grading standard for the statements composed by elementary school students. The study used the statements of the sentences made of words by pupils as training statements, and used the Latent Semantic Analysis method and the word semantics provided by Sinica Corpus to analyze the semantic weights of words and statements individually for the training statements, so as to create the grading standard. Finally, the study used the cosine function of Vector Space model to obtain all the similar statements to the target statement, so as to integrate the words of all the similar statements and the semantic weights of individual statements, and to obtain the grading score of target statement. The method of the study can assist the users to create the grading standard according to the overall writing degree, and to analyze the grading score of statements according to the semantic degree of similar statements. In addition, Halim and Khan [25] categorized academic journals based on data science methods using multiple significant bibliometric indicators. The dataset collected for 660 journals is preprocessed to fill-in the missing values and performed scaling. Three feature selection techniques are used to rank the 19 features (e.g., publisher, impact factor) and k-means and k-medoids are employed to obtain the optimum number of coherent groups in the dataset. Then, this study used k-NN (Nearest Neighbor) and Artificial Neural Network (ANN) to predict the category of an unknown journal. In addition, a descriptive analysis of the clusters formed is presented to gain insights about the four journal categories to improve academic journals classification.

2.3 Discussion about article writer behavior

Lin et al. [20] used Snippet-based Unsupervised approach to propose an emotion classification model. At first, the Google search engine was used to search for the fragments of training words and the emotional phrases of words were extracted for the adjectives and adverbs of articles. Then, the correlation between emotional phrases was calculated by using mutual information algorithm, and the less correlated phrases were used to create the phrases of opposite emotion. Finally, the polarity trend of emotional phrases (i.e. trend of positive and negative emotions) was calculated to obtain the emotion type of phrases. Li and Liu [7] proposed a clustering-based emotion analysis method. The study used the K-means Clustering Algorithm to divide the documents of document library into control groups of positive and negative emotions according to the semantic orientation of adjectives. The study used the TF-IDF algorithm to calculate the importance weights of target document and all documents in the document database. The study uses voting mechanism to calculate the importance of object document and two control groups, till the importance of target document to a control group approached to stableness; then, the emotion inclination of target document could be obtained.

Ichifuji et al. [18] used Morphological Analysis to analyze the compound words and double words of sentence subject, verb and object of intentionally destructive comments to create the intentionally destructive word stock. The comments were divided into general comments, unintentionally destructive and intentionally destructive comments. The Bayesian Filter methodology was used to analyze the double words and part of speech of comments to obtain the intentionally destructive comments. Finally, the characteristics of part of speech, compound words and double words of general comments were used to match intentionally destructive comments. After that, the probability of intentionally destructive comments being unintentionally destructive comments can be calculated to filter the unintentionally destructive comments implied in the intentionally destructive comments and to obtain the intentionally destructive comments more accurately. David et al. [6] tested 17 subjects to test the users' cognition of statements. The subjects were normal adults, the average age was 25, and they were right-handers. When the subjects saw the first statement in the article, the study used magnetoencephalogram to track the response of the subjects' brains to the correlation between this statement and the topic, and provided "correlated" and "uncorrelated" buttons for the subjects to choose. Then, the study observes the judgment time of the subjects' brains, the brain substance gain, and the subjects' cognition of the degree of correlation between statement and topic to analyze the cognitive data. The findings showed that the subjects cognized different meanings of the first statement, but if there were specifications or the article was presented structurally, the differences in the subjects' cognition of the statement were reduced apparently.

3.0 AN ARTICLE REORGANIZATION MODEL BASED ON EMOTION IMPLIED IN FORUM ARTICLES

The proposed "Article Reorganization Model based on the Emotion Implied in Forum Articles" analyzes the forum articles of virtual forums. After the word segmentation by the Chinese word segmentation system of Chinese Knowledge Information Processing Group (CKIP), the candidate event of forum articles can be obtained, the part of speech of candidate event, context and title correlation score is used to analyze the representative event of forum articles, and the similarity between the statements of representative event and the statements of training articles is worked out to analyze the emotion type of forum articles. Therefore, the subsequent “publisher article statement reorganization module” reorganizes the structure of text, consistent with the emotion to be expressed by the original article provider. The articles applicable to recomposition are obtained from the forum article review scores, and the semantically similar statements with emotion inconsistent with the original text are removed. The candidate filler sentences and combined sentences for the forum articles are analyzed, and multiple combination is implemented to reorganize the forum article structure. Finally, a forum article content with complete meaning is formed. The main process and architecture of this model can be divided into two major parts, which are Part1 "Article Expressed Emotion Determination Module" and Part2 "Publisher Article Statement Reorganization Module" in Fig. 3.1.

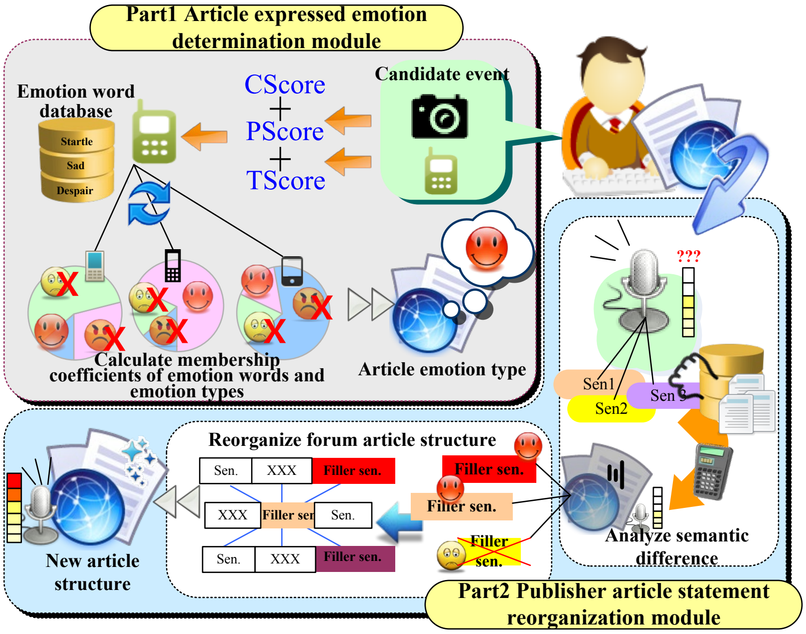

Part1 Article expressed emotion determination module
Emotion word database
Startle
Sad
Despair
CScore
PScore
TScore
Candidate event
Calculate membership coefficients of emotion words and emotion types
Article emotion type
???
Sen1
Sen2
Sen3
Analyze semantic difference
Reorganize forum article structure
Sen.
XXX
Filler sen.
Filler sen.
XXX
Filler sen.
Sen.
Filler sen.
Sen.
XXX
Filler sen.
Filler sen.
New article structure
Part2 Publisher article statement reorganization module

Fig. 3.1: The architecture of the proposed model

3.1 Article Expressed Emotion Determination Module

Step (A1)─Calculate context pairing score of article event and title weight score

According to the relationship between the statements before and after candidate event ACE$_i$ (i.e. statements between punctuation marks) and the event trigger words, the trigger words and candidate event ACE$_i$ have distribution rules. Therefore, this step calculates the distribution ratio of trigger words in the statements before and after target candidate event and the forum articles and works out the average, so as to obtain the context pairing mean score of target candidate event ACE$_i$. The larger the value is, the better the candidate event represents the article event. In addition, the title of forum article may represent the substance of article. Therefore, the title weighted score can be calculated according to the relationship between the forum article candidate event ACE$_i$ and title. Finally, the sum of the two scores is the context pairing score of candidate event and title weighted value TCScore(ACE$_i$), expressed as Equation (3.1).

$$\mathrm{TCScore}(\mathrm{ACE}_i)=\begin{bmatrix}\sum_{\text{all } j}\left(\dfrac{\mathrm{Fre}(\mathrm{CES}^{F}_{i,j})}{\mathrm{N}(\mathrm{CES}^{F}_{\bullet,\bullet})}+\dfrac{\mathrm{N}(\mathrm{CES}^{F}_{i,j})}{\mathrm{N}(\mathrm{ACE}_{\bullet})}\right)\\ \sum_{\text{all } k}\left(\dfrac{\mathrm{Fre}(\mathrm{CES}^{B}_{i,k})}{\mathrm{N}(\mathrm{CES}^{B}_{\bullet,\bullet})}+\dfrac{\mathrm{N}(\mathrm{CES}^{B}_{i,k})}{\mathrm{N}(\mathrm{ACE}_{\bullet})}\right)\\ \mathrm{Fre}(\mathrm{ACE}_i,\mathrm{Title})\\ \end{bmatrix}\cdot\begin{bmatrix}\dfrac{1}{\mathrm{N}(\mathrm{CES}^{F}_{i,\bullet})} & \dfrac{1}{\mathrm{N}(\mathrm{CES}^{B}_{i,\bullet})} & \dfrac{1}{\mathrm{Fre}(\mathrm{ACE}_i)}\end{bmatrix} \tag{3.1}$$

Step (A2)─Obtain part of speech correlation mean score of forum article event

As the part of speech of event and the part of speech of words before and after event have association rules, according to the part of speech established by Sinica Corpus, the proportion of part of speech of all words in all articles is calculated, and the part of speech distance parameter value Sdis(POS$_\bullet$) is marked (Table 3.1), so as to train the precision of part of speech distance. The part of speech correlation mean score of event is calculated as Equations (3.2) and (3.3). First, all candidate events are calculated Sdis(POS$_\bullet$) word sets Set(VEF) forward, and the occurrence number of target candidate event forward calculated words in word set Set(VEF) and articles to obtain the part of speech correlation probability of various words, and to calculate the part of speech correlation mean score of candidate event PScore(ACE$_i$).

Table 3.1: The summarization of category of speech all candidate events in training data

Category of part of speech	Occurrence number	Proportion	Part of speech distance
POS$_1$=Na (Common Nouns)	36	34.6%	Sdis(POS$_\bullet$)=1
POS$_2$=VE (Object Verbs)	16	15.4%	Sdis(POS$_\bullet$)=2
POS$_3$=NH (Pronoun)	16	15.4%	Sdis(POS$_\bullet$)=3

$$\mathrm{Set(VEF)}=\begin{bmatrix}\mathrm{VEF}_{1,1} & \mathrm{VEF}_{2,1}\cdots\mathrm{VEF}_{i,1}\\ \mathrm{VEF}_{1,2} & \mathrm{VEF}_{2,2}\cdots\mathrm{VEF}_{i,2}\\ \vdots & \vdots \quad\quad \vdots\\ \mathrm{VEF}_{1,n} & \mathrm{VEF}_{2,n}\cdots\mathrm{VEF}_{i,n}\end{bmatrix} \tag{3.2}$$

$$\mathrm{PScore(ACE_i)}=\frac{\sum_{\mathrm{all\ n}}\frac{\mathrm{Fre[Set(VEF_{i,n})]}}{\mathrm{N(VEF_{i,n})}}}{\beta(\mathrm{POS}_\bullet)\cdot\mathrm{N(ACE}_\bullet)\cdot\mathrm{Fre[Set(VEF_i)]}} \tag{3.3}$$

Step (A3)─Obtain representative event of forum articles

Considering different meanings and title importance of forum article types, this step adds up the part of speech correlation average of candidate event, context pairing score and title weight score and proportion score obtained by Step (A1) and Step (A2), the representative scores of various candidate events can be obtained. The candidate event with the maximum representative score is the representative event of article, expressed as Equation (3.4).

$$\begin{aligned}\mathrm{EScore(ACE_i)}&=\mathrm{PPScore(ACE_i)}\cdot\mathrm{PScore(ACE_i)}\\&+\mathrm{PTCScore(ACE_i)}\cdot\mathrm{TCScore(ACE_i)}\\&\mathrm{Where\,PPScore(ACE_i)}+\mathrm{PTCScore(ACE_i)}=1\end{aligned} \tag{3.4}$$

Step (A4)─Create membership coefficients of emotion words and emotion types

This step converts the 482 common English emotion words and corresponding emotion types constructed by Gregory and Daniel [12] into Chinese words, and divides the emotion types $\mathrm{S_w}$ into seven types, including Angry, Anxiety, Disgust, Fear, Happiness, Sadness and Surprise. The emotion statement set $\mathrm{SS\_Set_w}$ (Equation (3.5)) is created according to the emotion types corresponding to the existing emotion words (as shown in Table 3.2, represented by 3 types). Afterwards, the occurrence frequency of target emotion word in the emotion statement set $\mathrm{SS\_Set_w}$ is calculated referring to the method of Miao et al. [22], and the total number of emotion words of various emotion types in the training articles is calculated, namely, the coefficient of correlation $\mathrm{ReS(DE_d,S_w)}$ between target emotion word and various emotion types is calculated by Equation (3.6), and the coefficient of correlation is normalized, as shown in Equation (3.7), so as to know the membership coefficient of target emotion word and emotion type. The results are shown in Table 3.3, the larger the coefficient value is, the closer is the emotion word to the corresponding emotion type.

$$\mathrm{SS\_Set_w}=\mathrm{L_{p,b}}\ \text{where}\ \mathrm{LS[S_w,DE_d]}\ \text{exist in}\ \mathrm{L_{p,b}}\ \forall\mathrm{d} \tag{3.5}$$

$$\mathrm{ReS}'(\mathrm{DE_d,S_w})=\frac{\mathrm{N(DE_d}\cap\mathrm{SS\_Set_w)}}{\sum_{\mathrm{all\,d}}\mathrm{N[L_\bullet,LS(S_w,DE_d)]}}\times\log_2(\mathrm{N(DE_d}\cap\mathrm{SS\_Set_w})+1) \tag{3.6}$$

$$\mathrm{ReS(DE_d,S_w)}=\frac{\mathrm{ReS}'(\mathrm{DE_d,S_w})}{\sum_{\mathrm{all\,w}}\mathrm{ReS}'(\mathrm{DE_d,S_w})} \tag{3.7}$$

Table 3.2: Emotion types corresponding to the emotion words

Emotion types (English/Chinese)	English emotion words	Chinese emotion words
Angry/憤怒	Contempt	鄙視、輕視、藐視
	Violent	激烈、暴力
Anxiety/焦慮	Awkward	尷尬、笨拙、不熟練的
	Uneasy	不安、不自在、擔心
Disgust/厭惡	Villain	壞人、惡棍
	Stinking	惡臭、非常討厭

Table 3.3: Membership coefficient of emotion word and emotion type

Emotion word Emotion type	DE$_1$	DE$_2$	…	DE$_d$	…
S$_1$	ReS[DE$_1$，S$_1$]	ReS[DE$_2$，S$_1$]	…	ReS[DE$_d$，S$_1$]	…
…	…	…	…	…	…
S$_w$	ReS[DE$_1$，S$_w$]	ReS[DE$_2$，S$_w$]	…	ReS[DE$_d$，S$_w$]	…

Step (A5)─Calculate the similarity between representative statement of forum article and statements in all training articles

When the representative event of forum articles is obtained through the first three steps, this paper will train the statement vector set L_b^{ω} of articles and the set vector $ADS^{\omega}{}_q$ of target article representative statement ADS_q (i.e. statements with representative event), uses the cosine function of Vector Space model to calculate the similarity between the representative statement of target forum article and the statements in all training articles, and uses Equation (3.8) to judge the similarity $Sim(ADS_q, L_b)$ between the representative statement ADS_q and all statements.

$$
\begin{aligned}
& L_{p,b} = \left\{ L_{p,1}, L_{p,2}, L_{p,3}, \cdots, L_{p,b}, \cdots \right\} \\
& L_b^{\omega} = \left[w_1, w_2, \cdots, w_b \right]^T \\
& ADS^{\omega}{}_q = \left[w_1, w_2, \cdots, w_q \right]^T \\
& Sim(ADS_q, L_b) = \frac{ADS^{\omega}{}_q \cdot L^{\omega}{}_b}{|\, ADS^{\omega}{}_q \,| \cdot |\, L^{\omega}{}_b \,|}
\end{aligned}
\qquad (3.8)
$$

Step (A6)─Calculate emotion type probability value of similar statements

When the similarity $Sim(ADS_q, L_b)$ between forum article representative statement ADS_q and all statements is obtained, if the similarity $Sim(ADS_q, L_b)$ is greater than threshold $\omega(ADS_q, L_b)$ and close to 1, meaning the article statement and representative statement ADS_q have similar meanings. However, the statement is usually opposite to the original meaning due to privatives, so if the statement privative frequency is higher than the parameter value $\alpha(NW)$, the statement is dissimilar to the representative statement. In addition, this step extracts the emotion words and important words from similar statements to form the polarity item combination of representative statement. If the similar statements have no noun Na or place word Nc and emotion word, the polarity item combination of the similar statements is neglected, so as to keep the emotion word and important word correlation of the representative statement, expressed as Equation (3.9). The membership coefficient $ReS(DE_d, S_w)$ of the emotion words and emotion types contained in the polarity item is used as the membership coefficient $R[ADS\_DI_{q,y}, S_w]$ of the polarity item and emotion type, expressed as Equation (3.10). Finally, the probability value $SP[Fi(ADS_q, S_u), S_w]$ of all similar statements inclining to various emotion types is calculated by Equation (3.11). The results are shown in Table 3.4.

$$\begin{aligned}&\text{IF } \mathrm{Sim}(\mathrm{ADS}_q, \mathrm{L}_b) \geq \omega(\mathrm{ADS}_q, \mathrm{L}_b) \text{ and } \mathrm{N}(\mathrm{L}_b \cap \mathrm{Set(NW)}) < \alpha(\mathrm{NW}) \text{ and}\\ &\text{NANC exist in } \mathrm{L}_b \text{ and } \mathrm{DE}_d \text{ exist in } \mathrm{L}_b \text{ Then } \mathrm{Sim}(\mathrm{ADS}_q, \mathrm{L}_b) \in \mathrm{Sim\_Fi}(\mathrm{ADS}_q, \mathrm{S}_u)\\ &\text{and NANC}, \mathrm{DE}_d \in \mathrm{ADS\_DI}_q \ \forall_d\end{aligned} \tag{3.9}$$

$$\mathrm{R}[\mathrm{ADS\_DI}_{q,y}, \mathrm{S}_w] = \mathrm{ReS}(\mathrm{DE}_d, \mathrm{S}_w) \text{ where } \mathrm{DE}_d \text{ exist in } \mathrm{ADS\_DI}_{q,y} \ \forall \mathrm{d} \tag{3.10}$$

$$\mathrm{SP}[\mathrm{Fi}(\mathrm{ADS}_q,\mathrm{S}_u),\mathrm{S}_w] = \frac{\exp\left(\mathrm{I}[\mathrm{Fi}(\mathrm{ADS}_q,\mathrm{S}_u),\mathrm{TW})] + \mathrm{Sim\_Fi}(\mathrm{ADS}_q,\mathrm{S}_u) \times \dfrac{\sum_{\text{all y}} \mathrm{R}[\mathrm{ADS\_DI}_{q,y},\mathrm{S}_w]}{\mathrm{N}(\mathrm{ADS\_DI}_{q,y} \cap \mathrm{Fi}(\mathrm{ADS}_q,\mathrm{S}_u))}\right)}{\sum_{\text{all w}} \exp\left(\mathrm{I}[\mathrm{Fi}(\mathrm{ADS}_q,\mathrm{S}_u),\mathrm{TW})] + \mathrm{Sim\_Fi}(\mathrm{ADS}_q,\mathrm{S}_u) \times \dfrac{\sum_{\text{all y}} \mathrm{R}[\mathrm{ADS\_DI}_{q,y},\mathrm{S}_w]}{\mathrm{N}(\mathrm{ADS\_DI}_{q,y} \cap \mathrm{Fi}(\mathrm{ADS}_q,\mathrm{S}_u))}\right)} \tag{3.11}$$

$$\text{where } \mathrm{I}[\mathrm{Fi}(\mathrm{ADS}_q,\mathrm{S}_u),\mathrm{TW})] = \begin{Bmatrix} 0, \ \mathrm{N}(\mathrm{Fi}(\mathrm{ADS}_q,\mathrm{S}_u) \cap \mathrm{TW}) > 0 \\ 1, \text{otherwise} \end{Bmatrix} \text{ and}$$

$$\mathrm{ADS\_DI}_{q,y} \text{ exist in } \mathrm{Fi}(\mathrm{ADS}_q,\mathrm{S}_u) \ \forall \ \mathrm{y}$$

Table 3.4: The probability value of all similar statements inclining to various emotion types

Similar statement Emotion type	Fi(ADS$_q$，S$_1$)	Fi(ADS$_q$，S$_2$)	…	Fi(ADS$_q$，S$_u$)	…
S$_1$	SP[Fi(ADS$_q$，S$_1$)，S$_1$]	SP[Fi(ADS$_q$，S$_1$)，S$_1$]	…	SP[Fi(ADS$_q$，S$_u$)，S$_1$]	…
…	…	…	…	…	…
S$_w$	SP[Fi(ADS$_q$，S$_1$)，S$_w$]	SP[Fi(ADS$_q$，S$_2$)，S$_w$]	…	SP[Fi(ADS$_q$，S$_u$)，S$_w$]	…

Step (A7)─Obtain emotion type of forum article

The stable value $\mathrm{ST}[\mathrm{AL\_DI}_q,\mathrm{S}_w]$ of forum article representative statement for emotion type can be judged by Potts model (Equation (3.12)) according to the probability value $\mathrm{SP}[\mathrm{Fi}(\mathrm{ADS}_q,\mathrm{S}_u),\mathrm{S}_w]$ of similar statements and various types of emotion (as shown in Table 3.5). The closer the value is to 0, the closer is the article representative statement to the emotion type. However, the articles often contain several representative statements, and the forum articles often imply several emotion types, so this paper regards the emotion types of representative statements as the emotion types expressed by target forum article.

$$\begin{aligned}\mathrm{ST}[\mathrm{ADS}_q,\mathrm{S}_w] = &-\sum_{\text{all u}} \mathrm{SP}[\mathrm{Fi}(\mathrm{ADS}_q,\mathrm{S}_u),\mathrm{S}_w] \times \mathrm{I}[\mathrm{Fi}(\mathrm{ADS}_q,\mathrm{S}_u),\mathrm{TW})]\\ &-\sum_{\text{all u}} \mathrm{SP}[\mathrm{Fi}(\mathrm{ADS}_q,\mathrm{S}_u),\mathrm{S}_w] \times \mathrm{Sim\_Fi}(\mathrm{ADS}_q,\mathrm{S}_u) \times \mathrm{I}[\mathrm{Fi}(\mathrm{ADS}_q,\mathrm{S}_u),\mathrm{TW})]\\ &-\sum_{\text{all u}} - \mathrm{SP}[\mathrm{Fi}(\mathrm{ADS}_q,\mathrm{S}_u),\mathrm{S}_w] \times \log(\mathrm{SP}[\mathrm{Fi}(\mathrm{ADS}_q,\mathrm{S}_u),\mathrm{S}_w])\end{aligned} \tag{3.12}$$

Table 3.5: The stable value of representative statement for emotion type

Representative statement Emotion type	ADS$_1$	ADS$_2$	…	ADS$_q$	…
S$_1$	ST[ADS$_1$，S$_1$]	ST[ADS$_2$，S$_1$]	…	ST[ADS$_q$，S$_1$]	…
…	…	…	…	…	…
S$_w$	ST[ADS$_1$，S$_w$]	ST[ADS$_2$，S$_w$]	…	ST[ADS$_q$，S$_w$]	…
…	…	…	…	…	…

3.2 Publisher Article Statement Reorganization Module

Step (B1)─Obtain initial score of forum article

This step uses the surface feature "number of dissimilar words" of target forum article A_T as the basis of initial score. The repeated words of article are removed and the same words are regarded as one word class, so as to obtain the actual number of dissimilar words $N(FAL\_Set)$ in the target forum article. In addition, this step calculates the number of dissimilar words $N(LDW\_Set_p)$ in all forum articles of training article repository to match the similarity with the target forum article A_T in the subsequent steps (Equation (3.13)).

$$\begin{aligned} ADW\_Set &= \{AL_j \mid AL_j \, not\, exist\, in\, ADW\_Set\, \forall j\} \\ LDW\_Set_p &= \{L_{p,k} \mid L_{p,k} \, not\, exist\, in\, LDW\_Set_p\, \forall k\} \end{aligned} \tag{3.13}$$

Step (B2)─Calculate review score of target forum article

This step uses the shared words of target forum article and training article as the similarity $Sim(A_T, L_p)$ between the two articles, and calculates the standard deviation of similarity $Sd(A_T, L_\bullet)$, expressed as Equation (3.14), so that the review score can converge in the review range. Afterwards, the semantic difference $Sem(A_T, L_p)$ between the training article and target article is calculated by Equation (3.15) to obtain the difference in the shared words of training articles. Finally, the review score $PGrade$ of target forum article is calculated by Equation (3.16). If the review score $PGrade$ is lower than the threshold Th_PGrade, the semantic expression of the target forum article is relatively low.

$$\begin{aligned} Sim(A_T, L_p) &= N(ADW\_Set \cap LDW\_Set_p) \\ Sd(A_T, L_\bullet) &= \sqrt{\frac{\sum_{all\ p}\left(Sim(A_T, L_p) - \left(\frac{\sum_{all\ p} Sim(A_T, L_p)}{L_\bullet}\right)\right)^2}{L_\bullet - 1}} \end{aligned} \tag{3.14}$$

$$Sem(A_T, L_p) = \frac{\left(\frac{N(LDW\_Set_p) - \sum_{p \neq m} N(LDW\_Set_m)}{L_\bullet}\right)}{Sd(A_T, L_p)} \tag{3.15}$$

$$PGrade = \sum_{all\ p} Sim(A_T, L_p) \times Sem(A_T, L_p) \tag{3.16}$$

Step (B3)─Remove semantically similar statements

As Chinese statements often have sentences composed of different words but the same meaning, this step removes the statements with the same meaning in the forum articles to reorganize the important statements more accurately. This paper uses Chinese-English translator (Denisowski's CEDICT) to obtain the English words $ASL_{m,i}$ of various words in the target statement of forum article, and then uses WordNet to obtain the English words of different parts of speech to extend the defined word set $ASLE\_Set_m$, and calculates the correlation grade $Re[ASL_{m,i}, ASL_{k,i}]$ of English words of all statements, expressed as Equation (3.17). Finally, the degrees of correlation between all target statements and the words of other statements are summed up, i.e. the ambiguous similarity $amp[AS_m, AS_k]$ between target statement and other statements, expressed as Equation (3.18), and the statement with maximum ambiguous similarity is used as the ambiguous sentence of target statement. This paper removes the statements with the least parts of speech in target statement and ambiguous sentence to obtain the important statements of forum articles.

$$\begin{aligned}&\mathrm{CL\_Set}_i = \mathrm{ASLE\_Set}_{m,i} \cap \mathrm{ASLE\_Set}_{k,i}\ \forall i\\&\mathrm{Re}[\mathrm{ASL}_{m,i},\mathrm{ASL}_{k,i}] = \mathrm{Max}(\text{-}\log(\frac{\mathrm{N}(\mathrm{CL\_Set}_i[\mathrm{EL}_u])}{\mathrm{ASL}_{\bullet,i}}))\ \forall u\end{aligned} \tag{3.17}$$

$$\mathrm{amp}[\mathrm{AS}_m,\mathrm{AS}_k] = \sum_{\text{all i}} \mathrm{Re}[\mathrm{ASL}_{m,i},\mathrm{ASL}_{k,i}] \tag{3.18}$$

Step (B4)─Obtain candidate filler sentences of important statements and representative statements

The important statements and representative statements (statements with representative event) are obtained in previous steps, this step forms extended statement set $\mathrm{Ex\_Set}$ (Equation (3.19)) according to the order of statements in the forum articles, and works out the similarity between the statements in the extended statement set $\mathrm{Ex\_Set}$ and the statements $\mathrm{L}_{\bullet\bullet}$ in training article repository (Equation (3.20)). The statements before and after the statement with maximum similarity are used as candidate filler sentences. In addition, in order to make the rewritten forum article consistent with the emotion to be expressed by the original forum article, this step filters the candidate filler sentences inconsistent with the original emotion of forum article, expressed as Equation (3.21).

$$\mathrm{Ex\_Set} = \{\mathrm{AS}_1,\mathrm{AS}_2,\mathrm{AS}_3,\cdots,\mathrm{AS}_m \mid \mathrm{Max}(\mathrm{amp}[\mathrm{AS}_m,\mathrm{AS}_k])\,\mathrm{and}\,\mathrm{Max}(\mathrm{ASL}_{m,\bullet})\} \tag{3.19}$$

$$\mathrm{Sim}(\mathrm{Ex\_Set}[\mathrm{IS}_g],\mathrm{L}_b) = \frac{\mathrm{Ex\_Set}[\mathrm{IS}^{\omega}{}_g]\cdot \mathrm{L}^{\omega}{}_b}{|\,\mathrm{Ex\_Set}[\mathrm{IS}^{\omega}{}_g]\,|\cdot|\,\mathrm{L}^{\omega}{}_b\,|} \tag{3.20}$$

$$\mathrm{Con\_Set} = \{\mathrm{L}_{b+1},\mathrm{L}_{b-1} \mid \mathrm{Max}(\mathrm{Sim}(\mathrm{Ex\_Set}[\mathrm{IS}_g],\mathrm{L}_b))\ \mathrm{and}\ \mathrm{LS}_{p,b,d} \in \mathrm{AS\_Set}_v\}\forall g, w \tag{3.21}$$

Step (B5)─Obtain multiple combined sentences of candidate filler sentences

As the combination of candidate filler sentence $\mathrm{Con\_Set}[\mathrm{CS}_q]$ and important statement and representative statement may result in incoherent sentence, this step compares the keywords of candidate filler sentence with the statements of all articles in training article repository, so as to obtain the filler word $\mathrm{LF}_{p,b,x}$ of candidate filler sentence (word between Keywords), to collect the multiple combined sentences $\mathrm{CS\_MUS}_{q,z}$ of candidate filler sentence, expressed as Equation (3.22).

$$\mathrm{CS\_MUS}_{q,z} = \left\{\begin{matrix}\mathrm{LF}_{p,b,x} \mid \mathrm{Con\_Set}[\mathrm{CS}_{q,r}]\,\text{exist in}\ \mathrm{L}_{p,b}\\ \text{and}\ \mathrm{Con\_Set}[\mathrm{CS}_{q,r}]\,\text{not exist in}\ \mathrm{LF}_{p,b,x}\end{matrix}\right\}\forall r, x \tag{3.22}$$

Step (B6)─Reorganize forum article structure

According to the multiple combined sentences of candidate filler sentences obtained in previous step, this step combines the multiple combined sentences $\mathrm{CS\_MUS}_{q,z}$ of candidate filler sentences, representative statement and important statement of target forum article according to the order of statements, expressed as Equation (3.23), so as to reorganize the statements of the forum article fit for reorganization. In addition, as the candidate filler sentences have multiple combined sentences, the reorganized forum article will have several different statement structures. In order to obtain the forum article of complete structure, this step obtains the review score $\mathrm{Rec\_Grade}$ of the reorganized forum article through Step (B1) and Step (B2) (as shown in Table 3.6), and the statement structure with the highest review score is used as the reorganized content of target forum article.

$$\begin{aligned}\mathrm{Rec\_A}_n = &\begin{bmatrix}\mathrm{CS\_MUS}_{1,1}\\ \mathrm{CS\_MUS}_{1,2}\\ \vdots\\ \mathrm{CS\_MUS}_{1,c}\end{bmatrix} \rightarrow \begin{bmatrix}\mathrm{CS\_MUS}_{2,1}\\ \mathrm{CS\_MUS}_{2,2}\\ \vdots\\ \mathrm{CS\_MUS}_{2,a}\end{bmatrix} \rightarrow \begin{bmatrix}\mathrm{CS\_MUS}_{3,1}\\ \mathrm{CS\_MUS}_{3,2}\\ \vdots\\ \mathrm{CS\_MUS}_{3,v}\end{bmatrix} \rightarrow\cdots\rightarrow \begin{bmatrix}\mathrm{CS\_MUS}_{q,1}\\ \mathrm{CS\_MUS}_{q,2}\\ \vdots\\ \mathrm{CS\_MUS}_{q,z}\end{bmatrix} \rightarrow\cdots\\ &\text{where}\ \mathrm{Max}(\mathrm{Rec\_Grade}_n)\end{aligned} \tag{3.23}$$

Table 3.6: The summarization of article statement structure and review score

Article statement structure	Article statement structure content	Review score
Rec_A$_1$	$CS\_MUS_{1,1} \rightarrow CS\_MUS_{2,2} \rightarrow \cdots \rightarrow CS\_MUS_{q,3}$	Rec_Grade$_1$
Rec_A$_2$	$CS\_MUS_{1,1} \rightarrow CS\_MUS_{2,3} \rightarrow \cdots \rightarrow CS\_MUS_{q,4}$	Rec_Grade$_2$
…	…	…
Rec_A$_{14}$	$CS\_MUS_{1,5} \rightarrow CS\_MUS_{2,1} \rightarrow \cdots \rightarrow CS\_MUS_{q,2}$	Rec_Grade$_2$
…	…	…
Rec_A$_n$	$CS\_MUS_{1,c} \rightarrow CS\_MUS_{2,a} \rightarrow \cdots \rightarrow CS\_MUS_{q,v}$	Rec_Grade$_n$

4.0 A WEB-BASED ARTICLE REORGANIZATION SYSTEM

According to the methodology proposed in Chapter 3, this paper develops a Web-based Article Reorganization system based on the emotion implied in forum articles to confirm the model feasibility. The system users are divided into common user and system administrator, and they are authorized with different functions usage. In addition, the system administrator collects training articles according to forum articles (e.g. Mobile01 forum) as the basis of subsequent analysis (as shown in Fig. 4.1 and Fig. 4.2).

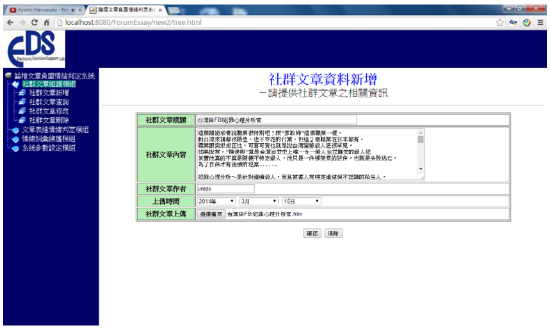
Fig. 4.1: Mobile01 forum article (1)

Fig. 4.2: Mobile01 forum article (2)

Upload forum Article

- ➢ Forum Article Uploading Function

When the system user executes Forum Article Uploading function and enters title "Taiwan and …" and content "this one title …", and enters the author "undio", and defines the user-defined path, e.g. "D:\data", the forum article file can be uploaded (as shown in Fig. 4.3). The system executes data preprocessing like word segmentation. The results "this(Nep) one(Nf) title(Na)…" are obtained after word segmentation, and the results are stored and maintained in the database (as shown in Fig. 4.4).

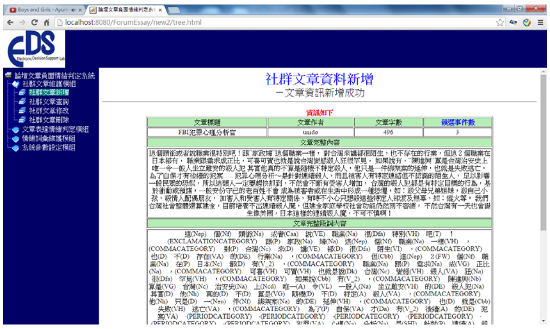


localhost:8080/ForumEssay/new2/tree.html

社群文章資料新增
—請提供社群文章之相關資訊

社群文章標題	台灣與FBI犯罪心理分析官
社群文章內容	這個頭銜或者說職業很特別吧！跟"家政婦"這個職業一樣， 對台灣來講都很陌生，也不存在的行業，但這2個職業在日本都有， 職業跟需求成正比，可喜可賀也就是說台灣變態殺人狂很罕見， 如果說有，"陳進興"算是台灣治安史上唯一令一般人坐立難安的殺人犯 其實他真的不算是隨機不特定殺人，他只是一件綁架案的延伸，也就是失敗逃亡， 為了自保才有後續的犯案…… 犯罪心理分析~是針對連續殺人，而且被害人有特定連結但不認識的陌生人，
社群文章作者	unido
上傳時間	2014年 3月 15日
社群文章上傳	選擇檔案 台灣與FBI犯罪心理分析官.htm

Fig. 4.3: Input forum article data



localhost:8080/ForumEssay/new2/tree.html

社群文章資料新增
—文章資訊新增成功

資訊如下

文章標題	文章作者	文章字數	候選事件數
FBI犯罪心理分析官	unido	496	3

文章完整內容

這個頭銜或者說職業很特別吧！跟"家政婦"這個職業一樣，對台灣來講都很陌生，也不存在的行業，但這2個職業在日本都有，職業跟需求成正比，可喜可賀也就是說台灣變態殺人狂很罕見，如果說有，"陳進興"算是台灣治安史上唯一令一般人坐立難安的殺人犯 其實他真的不算是隨機不特定殺人，他只是一件綁架案的延伸，也就是失敗逃亡，為了自保才有後續的犯案…… 犯罪心理分析~是針對連續殺人，而且被害人有特定連結但不認識的陌生人，足以影響一般民眾的恐慌，所以這類人一定要趕快抓到，不然會不斷有受害人增加，台灣的殺人犯都是有特定目標的行為，基於衝動或預謀，一般安份守己的老百姓不會成為被害者或在生活中形成一種恐懼，如：殺父母兄弟姐妹，殺自己小孩，殺情人配偶朋友，加害人和受害人有特定關係，有時不小心只想殺這些特定人卻波及無辜，如：縱火等。我們台灣社會整體還算建全，目前培養不出連續殺人魔，但是家庭學校社會功能仍然刻不容緩，不然台灣有一天也會跟美國，日本這樣的連續殺人魔，不可不慎啊！

文章完整斷詞內容

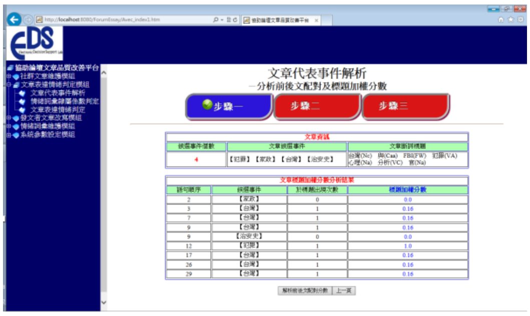
這(Nep) 個(Nf) 頭銜(Na) 或者(Caa) 說(VE) 職業(Na) 很(Dfa) 特別(VH) 吧(T) ！(EXCLAMATIONCATEGORY) 跟(P) 家政(Na) 婦(Na) 這(Nep) 個(Nf) 職業(Na) 一樣(VH) ，(COMMACATEGORY) 對(P) 台灣(Nc) 來(D) 講(VE) 都(D) 很(Dfa) 陌生(VI) ，(COMMACATEGORY) 也(D) 不(D) 存在(VA) 的(DE) 行業(Na) ，(COMMACATEGORY) 但(Cbb) 這(Nep) 2(FW) 個(Nf) 職業(Na) 在(P) 日本(Nc) 都(D) 有(V_2) ，(COMMACATEGORY) 職業(Na) 跟(P) 需求(Na) 成(VG) 正比(Na) ，(COMMACATEGORY) 可喜(VH) 可賀(VH) 也就是說(Dk) 台灣(Nc) 變態(VH) 殺人(VA) 狂(Na) 很(Dfa) 罕見(VH) ，(COMMACATEGORY) 如果說(Cbb) 有(V_2) ，(COMMACATEGORY) 陳進興(Nb) 算是(VG) 台灣(Nc) 治安史(Na) 上(Ncd) 唯一(A) 令(VL) 一般人(Na) 坐立難安(VH) 的(DE) 殺人犯(Na) 其實(D) 他(Nh) 真的(D) 不(D) 算是(VG) 隨機(D) 不(D) 特定(A) 殺人(VA) ，(COMMACATEGORY) 他(Nh) 只是(D) 一(Neu) 件(Nf) 綁架案(Na) 的(DE) 延伸(VH) ，(COMMACATEGORY) 也(D) 就是(Cbb) 失敗(VH) 逃亡(VA) ，(COMMACATEGORY) 為了(P) 自保(VA) 才(Da) 有(V_2) 後續(A) 的(DE) 犯案(VA) .(PERIODCATEGORY) .(PERIODCATEGORY) .(PERIODCATEGORY) .(PERIODCATEGORY) .(PERIODCATEGORY) .(PERIODCATEGORY) 犯罪(VA) 心理(Na) 分析(Na) 是(SHI) 針對(P) 連續(A)

Fig. 4.4: forum article data preprocessing

Article-expressed Emotion Inference

➢ Article Representative Event Analysis Function

The system administrator executes article representative event analysis function for the article titled "Taiwan and FBI.", the system captures the candidate event of target article automatically. When the candidate event "Taiwan" occurs in the title "1" time, the calculated title weighted score is "0.16"(as shown in Fig. 4.5). Afterwards, the context pairing score "0.4" of candidate event can be obtained according to the distribution value of contextual statement words (as shown in Fig. 4.6). The system calculates the distribution average "0.24" of all candidate event preambles according to the part of speech and context words (as shown in Fig. 4.7). When three scores are analyzed, the system normalizes three total scores of candidate event to know that the representative score "0.459" of candidate event "crime" is maximum value, so the "crime" is the representative event of article (as shown in Fig. 4.8).



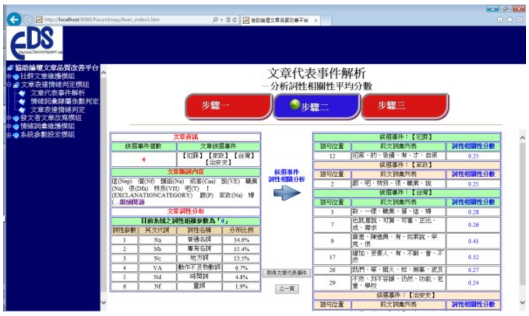
文章代表事件解析
—分析前後文配對及標題加權分數

文章資訊

候選事件個數	文章候選事件	文章斷詞標題
4	【犯罪】【家政】【台灣】【治安史】	台灣(Nc) 與(Caa) FBI(FW) 犯罪(VA) 心理(Na) 分析(VC) 書(Na)

文章標題加權分數分析結果

語句順序	候選事件	於標題出現次數	標題加權分數
2	【家政】	0	0.0
3	【台灣】	1	0.16
7	【台灣】	1	0.16
9	【台灣】	1	0.16
9	【治安史】	0	0.0
12	【犯罪】	1	1.0
17	【台灣】	1	0.16
26	【台灣】	1	0.16
29	【台灣】	1	0.16



文章代表事件解析
—分析詞性相關性平均分數

文章資訊

候選事件個數	文章候選事件
4	【犯罪】【家政】【台灣】【治安史】

文章斷詞內容

這(Nep) 個(Nf) 類似(Na) 或者(Caa) 說(VE) 職業(Na) 很(Dfa) 特別(VH) 吧(T) !(EXCLANATIONCATEGORY) 跟(P) 家政(Na) 婦(…繼續閱讀

文章詞性分布

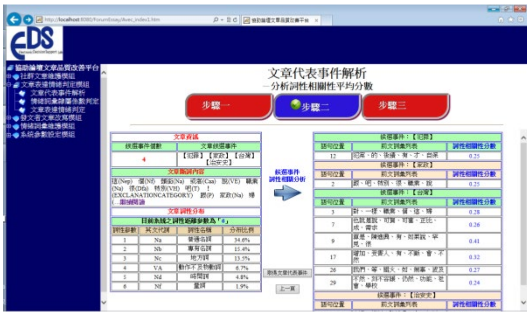
目前系統之詞性距離參數為「6」

詞性參數	英文代碼	詞性名稱	分布比例
1	Na	普通名詞	34.6%
2	Nb	專有名詞	15.4%
3	Nc	地方詞	13.5%
4	VA	動作不及物動詞	6.7%
5	Nd	時間詞	4.8%
6	Nf	量詞	1.9%

候選事件：【犯罪】		
語句位置	前文詞彙列表	詞性相關性分數
12	犯案、的、後續、有、才、自保	0.25
候選事件：【家政】		
語句位置	前文詞彙列表	詞性相關性分數
2	跟、吧、特別、很、職業、說	0.25
候選事件：【台灣】		
語句位置	前文詞彙列表	詞性相關性分數
3	對、一樣、職業、個、這、婦	0.28
7	也就是說、可算、可靠、正比、成、需求	0.26
9	竟是、陳進興、有、如果說、罕見、很	0.41
17	增加、受害人、有、不斷、會、不然	0.32
26	我們、等、縱火、如、謀事、波及	0.27
29	不然、刻不容緩、仍然、功能、社會、學校	0.24
候選事件：【治安史】		
語句位置	前文詞彙列表	詞性相關性分數

Fig. 4.5: Candidate event weighted value

Fig. 4.6: Distribution value of contextual statement words



文章代表事件解析
—分析詞性相關性平均分數

文章資訊

候選事件個數	文章候選事件
4	【犯罪】【家政】【台灣】【治安史】

文章斷詞內容

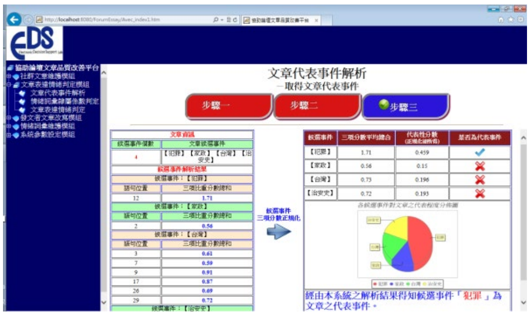
這(Nep) 個(Nf) 類似(Na) 或者(Caa) 說(VE) 職業(Na) 很(Dfa) 特別(VH) 吧(T) !(EXCLANATIONCATEGORY) 跟(P) 家政(Na) 婦(…繼續閱讀

文章詞性分布

目前系統之詞性距離參數為「6」

詞性參數	英文代碼	詞性名稱	分布比例
1	Na	普通名詞	34.6%
2	Nb	專有名詞	15.4%
3	Nc	地方詞	13.5%
4	VA	動作不及物動詞	6.7%
5	Nd	時間詞	4.8%
6	Nf	量詞	1.9%

候選事件：【犯罪】		
語句位置	前文詞彙列表	詞性相關性分數
12	犯案、的、後續、有、才、自保	0.25
候選事件：【家政】		
語句位置	前文詞彙列表	詞性相關性分數
2	跟、吧、特別、很、職業、說	0.25
候選事件：【台灣】		
語句位置	前文詞彙列表	詞性相關性分數
3	對、一樣、職業、個、這、婦	0.28
7	也就是說、可算、可靠、正比、成、需求	0.26
9	竟是、陳進興、有、如果說、罕見、很	0.41
17	增加、受害人、有、不斷、會、不然	0.32
26	我們、等、縱火、如、謀事、波及	0.27
29	不然、刻不容緩、仍然、功能、社會、學校	0.24
候選事件：【治安史】		
語句位置	前文詞彙列表	詞性相關性分數



文章代表事件解析
—取得文章代表事件

文章資訊

候選事件個數	文章候選事件
4	【犯罪】【家政】【台灣】【治安史】

候選事件解析結果

候選事件：【犯罪】	
語句位置	三項比重分數總和
12	1.71
候選事件：【家政】	
語句位置	三項比重分數總和
2	0.56
候選事件：【台灣】	
語句位置	三項比重分數總和
3	0.61
7	0.59
9	0.91
17	0.87
26	0.69
29	0.72
候選事件：【治安史】	

候選事件	三項分數平均總合	代表性分數 (正規化後所得)	是否為代表事件
【犯罪】	1.71	0.459	✓
【家政】	0.56	0.15	✗
【台灣】	0.73	0.196	✗
【治安史】	0.72	0.193	✗

各候選事件對文章之代表程度分佈圖

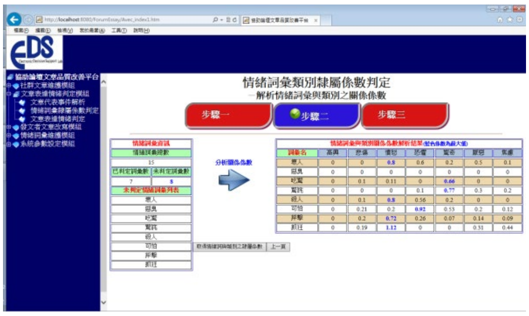
經由本系統之解析結果得知候選事件「犯罪」為文章之代表事件。

Fig. 4.7: Representative event of article determination (1)

Fig. 4.8: Representative event of article determination (2)

➢ Emotion Word Membership Coefficient Determination Function

AS the system administrator executes the emotion word membership coefficient determination function, the number of words of determined emotion type membership coefficient is "7", the number of undetermined words is "8", the correlation coefficient of emotion word "amaze" in emotion types "fear", "surprise", "disgust" and "anxiety" is calculated according to the number of words of emotion type and the frequency number of words in emotion type statements, which is "0.1", "0.77", "0.3" and "0.2" respectively. The correlation coefficient of the other 3 emotion types is "0" (as shown in Fig. 4.9). Finally, the system normalizes the correlation coefficient of words to obtain the maximum type membership coefficient "0.56" of emotion word "amaze", the corresponding emotion type is "surprise". The administrator can click "Review" button to know the membership distribution of word "amaze" in various emotion types (as shown in Fig. 4.10).



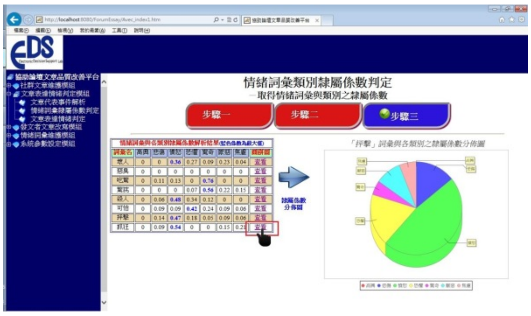
情緒詞彙類別隸屬係數判定
—解析情緒詞彙與類別之關係係數

情緒詞彙資訊	
情緒詞彙總數	
15	
已判定詞彙數	未判定詞彙數
7	8
未判定情緒詞彙列表	
嚇人	
惡臭	
吃驚	
驚訝	
殺人	
可怕	
抨擊	
抓狂	

情緒詞彙與類別關係係數解析結果(藍色係數為最大值)

詞彙名	高興	悲傷	憤怒	恐懼	驚奇	厭惡	焦慮
嚇人	0	0	0.8	0.6	0.2	0.5	0.1
惡臭	0	0	0	0	0	0	0
吃驚	0	0.1	0.11	0	0.66	0	0
驚訝	0	0	0	0.1	0.77	0.3	0.2
殺人	0	0.1	0.8	0.56	0.2	0	0
可怕	0	0.21	0.2	0.92	0.53	0.2	0.12
抨擊	0	0.2	0.72	0.26	0.07	0.14	0.09
抓狂	0	0.19	1.12	0	0	0.31	0.44

Fig. 4.9: Calculate correlation coefficients of emotion words



情緒詞彙類別隸屬係數判定
—取得情緒詞彙與類別之隸屬係數

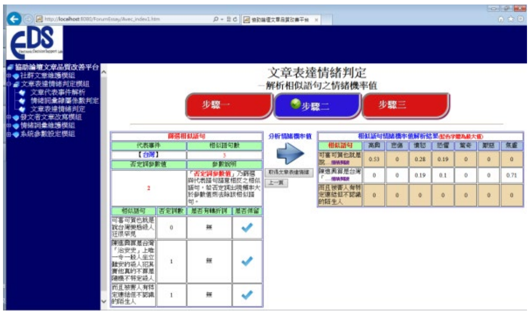
情緒詞彙與各類別隸屬係數解析結果(藍色係數為最大值)

詞彙名	高興	悲傷	憤怒	恐懼	驚奇	厭惡	焦慮	隸屬圖
嚇人	0	0	0.36	0.27	0.09	0.23	0.04	查看
惡臭	0	0	0	0	0	0	0	查看
吃驚	0	0.11	0.13	0	0.76	0	0	查看
驚訝	0	0	0	0.07	0.56	0.22	0.15	查看
殺人	0	0.06	0.48	0.34	0.12	0	0	查看
可怕	0	0.09	0.09	0.42	0.24	0.09	0.06	查看
抨擊	0	0.14	0.47	0.18	0.05	0.09	0.06	查看
抓狂	0	0.09	0.54	0	0	0.15	0.21	查看

Fig. 4.10: Calculate membership coefficients of emotion words

➢ Article-Expressed Emotion Determination Function

After the system administrator executes the article-expressed emotion determination function, the system analyzes the representative event "Taiwan" in the article, the system uses the cosine function of Vector Space model to calculate the similarity of representative statement. Afterwards, the system judges whether the similar statements have disjunctive words and privatives, and analyzes the emotion probability value of statement in emotion types according to similar statements (as shown in Fig. 4.11). Finally, the system calculates the emotional stability value "-0.726" of "fear" emotion type, "-0.202" of "surprise" and "-0.152" of "anxiety" according to the similar statements corresponding to representative statement, and stable value "-0.152" closest to 0 is the emotion types "anxiety" and "happiness" corresponding to representative statement (as shown in Fig. 4.12).



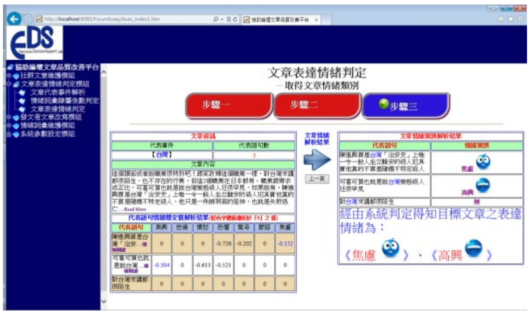
文章表達情緒判定
—解析相似語句之情緒機率值

篩選相似語句			
代表事件		相似語句數	
【台灣】		3	
否定詞參數值		參數說明	
2		「否定詞參數值」乃篩選與代表語句語意相近之相似語句，若否定詞出現頻率大於參數值則去除該相似語句。	
相似語句	否定詞數	是否有轉折詞	是否保留
可喜可賀也就是說台灣變態殺人狂很罕見	0	無	✓
陳進興算是台灣「治安史」上唯一令一般人坐立難安的殺人犯其實他真的不算是隨機不特定殺人	1	無	✓
而且被害人有特定連結但不認識的陌生人	1	無	✓

相似語句情緒機率值解析結果(藍色字體為最大值)

相似語句	高興	悲傷	憤怒	恐懼	驚奇	厭惡	焦慮
可喜可賀也就是說…情緒判定	0.53	0	0.28	0.19	0	0	0
陳進興算是台灣「…情緒判定	0	0	0.19	0.1	0	0	0.71
而且被害人有特定連結但不認識的陌生人	0	0	0	0	0	0	0

Fig. 4.11: Analyzes the emotion probability value of similar statements



文章表達情緒判定
—取得文章情緒類別

文章資訊	
代表事件	代表語句數
【台灣】	3
文章內容	

這個頭銜或者說職業很特別吧！跟家政婦這個職業一樣，對台灣來講都很陌生，也不存在的行業，但這2個職業在日本都有，職業跟需求成正比，可喜可賀也就是說台灣變態殺人狂很罕見，如果說有，陳進興算是台灣「治安史」上唯一令一般人坐立難安的殺人犯其實他真的不算是隨機不特定殺人，也只是一件綁架案的延伸，也就是失敗逃亡…Read More

代表語句情緒穩定值解析結果(藍色字體為最接近「0」之值)

代表語句	高興	悲傷	憤怒	恐懼	驚奇	厭惡	焦慮
陳進興算是台灣「治安…情緒判定	0	0	0	-0.726	-0.202	0	-0.152
可喜可賀也就是說台灣…情緒判定	-0.304	0	-0.613	-0.521	0	0	0
對台灣來講都很陌生	0	0	0	0	0	0	0

文章情緒類別解析結果

代表語句	情緒類別
陳進興算是台灣「治安史」上唯一令一般人坐立難安的殺人犯其實他真的不算是隨機不特定殺人	焦慮
可喜可賀也就是說台灣變態殺人狂很罕見	高興
對台灣來講都很陌生	無

經由系統判定得知目標文章之表達情緒為：

《焦慮》、《高興》

Fig. 4.12: Obtain emotion types of forum article

Publisher Article Statement Reorganization

➢ Article Review Score Determination Function

The system administrator executes article review score determination function, this paper takes a long article as an example, and selects an article titled "Taiwanese players even…", the system analyzes "172" different words in the article, and calculates the marking standard deviation "10" according to the number of dissimilar words (as shown in Fig. 4.13). Afterwards, the system calculates the semantic difference "-12.7" between the target article and training article content "recently, relatives ask…" according to the marking standard deviation, the number of common words is "29", and the semantic difference from the training article content "he says you don't… if you are unhappy" is "12.7", the number of common words is "49". Finally, the system calculates the target article review score "264" according to the semantic difference and the number of common words (as shown in Fig. 4.14).

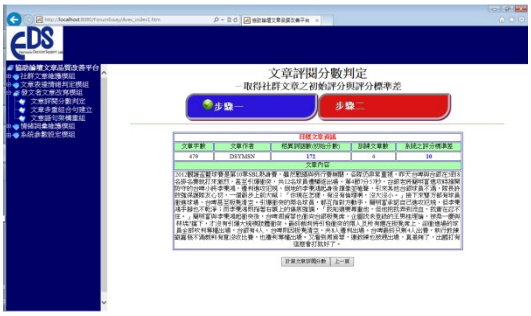


Fig. 4.13: Article review score determination (1)

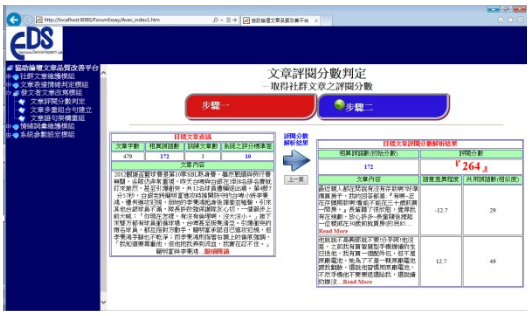


Fig. 4.14: Article review score determination (2)

➢ Article Multiple Combined Sentences Construction Function

After the system administrator executes article multiple combined sentences construction function, the system analyzes the intersection word "Snowfall" of English word extended definition set of target statement and other statement "weather of New York where I live..", the word occurs 3 times in the article, and the number of English pronouns in the article is 21, the system obtains the correlation grade "0.65" of target statement and important statement according to the aforesaid data analysis (as shown in Fig. 4.15). Afterwards, the system removes the semantic correlation grade "0.85" and statement content "but there is remaining snow freezing" according to the threshold condition (as shown in Fig. 4.16). The system matches the similarity of important statement "but there was a fall of snow this week", the similarity of statement "but there was a heavy fall of snow last week" is "0.707", the statement similarity is maximum value. Therefore, the system identifies the statement as candidate filler sentence of important statement "but there was a fall of snow this week" (as shown in Fig. 4.17). Afterwards, the system matches the keywords "weather, New York, stable" of candidate filler sentence "where I live" with training article library for coincident statements, five combined sentences of statements "the first cause…" and "weather of New York.." with Keywords are formed into multiple combined sentences of candidate filler sentence (as shown in Fig. 4.18).

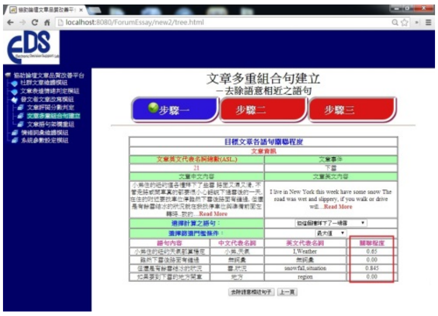


Fig. 4.15: Calculate correlation score among all statements

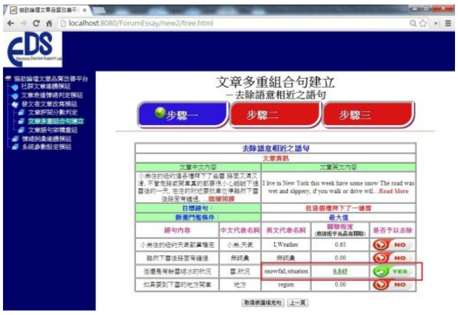


Fig. 4.16: Removes the similar semantic sentences

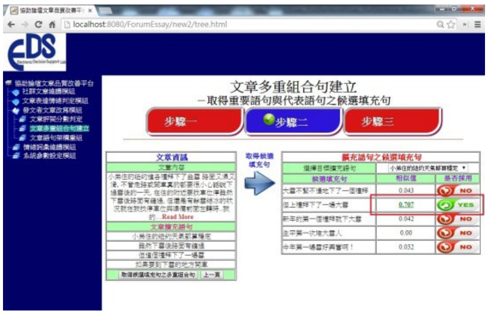


Fig. 4.17: Obtaine candidate filler sentences

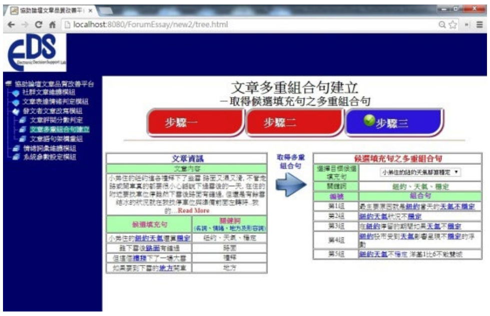


Fig. 4.18: Obtaine multiple combined sentences of candidate filler sentence

➢ Article Statement Structure Reorganization Function

After the article multiple combined sentences are constructed, the system obtains four statements to be replaced in the target article, there are another five combined sentences containing "but there was a fall of snow this week", there is another one combined sentence with "to drive where it snows", the article statements have "12" different statement structures. When the first combination is selected, the system replaces "but there was a fall of snow this week" with "but there was a heavy snow last week", and replaces "to drive where it snows" with "to drive when it's snowing", and the new article structure is displayed in the interface (as shown in Fig. 4.19). The second new article structure content is shown in Fig. 4.20. Afterwards, the system administrator can execute the final step, and click the article with the maximum review score, the system calculates the similarity standard deviation "1138.01" of the first article structure, and the fourth group has the maximum review score "1183.62" among all statement combinations (as shown in Fig. 4.21). Finally, the system displays the statement structure of the fourth group (as shown in Fig. 4.22).

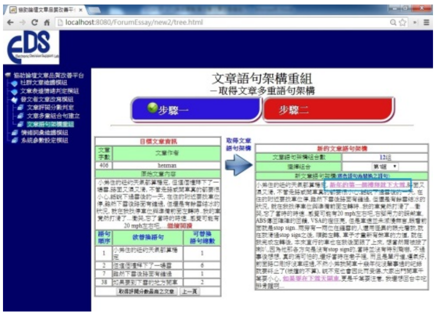


Fig. 4.19: Select the first new article structure

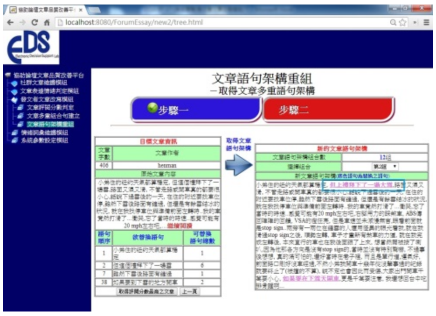
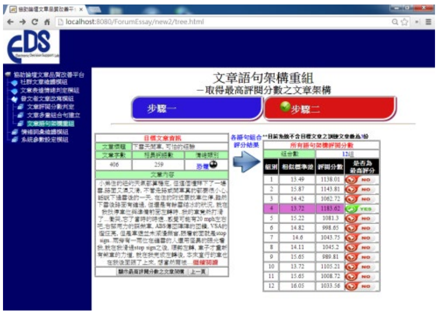


Fig. 4.20: Select the second new article structure



組別	相似標準差	評閱分數	是否為最高評分
1	13.49	1138.01	NO
2	15.87	1143.81	NO
3	14.42	1062.72	NO
4	13.72	1183.62	YES
5	15.22	1081.3	NO
6	14.82	998.65	NO
7	14.6	1043.75	NO
8	14.11	1045.2	NO
9	15.65	989.81	NO
10	13.72	1105.21	NO
11	15.65	1008.72	NO
12	16.05	1033.56	NO

Fig. 4.21: Obtain statement combinations based on the maximum review score

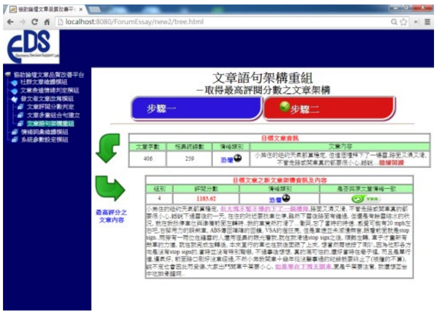

localhost:8080/ForumEssay/new2/tree.html
EDS
文章語句架構重組
一取得最高評閱分數之文章架構
步驟一
步驟二
目標文章資訊
406
259
1183.62
YES

Fig. 4.22: Display content of the obtained statement structure

5.0 CASE STUDY

The validation of "article-expressed emotion determination" can be divided into "comparative study" and "system performance evaluation"; in the validation of "publisher article statement reorganization", this paper uses the articles of "Mobile01 forum" as the samples of validation and training articles to validate the system performance. In addition, this paper uses recall rate, accuracy rate and F-measure as validation indexes.

5.1 Collection and construction of validation data

In the "article-expressed emotion determination", the validation data of comparative study are "PIXNET", and the system performance evaluation selects "Yahoo News" as source of validation data. In addition, in the "publisher article statement reorganization", this paper selects "Mobile01 forum" as the sample of validation and test data to validate the feasibility of this system. This paper collects the articles published by bloggers randomly in "PIXNET" as training articles, the short articles are screened in the course of collection to firm the training articles. "Yahoo News" contains a lot of news articles, and each news can be voted on the impressions. Therefore, this paper collects the "news article" data voted by over 30 person-times in various domains randomly. Finally, this paper collects training articles randomly from the top five topics of "Mobile01 forum" preambles, the actual articles are shown in Fig. 4.1. The short articles are screened during collection to construct the "candidate filler sentence set" of complete statements to obtain more accurate inference and validation results.

5.2 Validation of article-expressed emotion determination

The "article-expressed emotion determination" is divided into "comparative study" and "system performance evaluation" to validate the system performance. The verification models are detailed below, and the validation design, validation performance and results are described.

(A-1) Description of verification models compared with other studies

This paper collects 1300 validation data randomly from "PIXNET" blog (e.g. 2 actual articles in Table 5.1) referring to the validation scenario and validation data source of Tung and Lu [4], and collects additional 500 training articles, the actual emotion types of 1300 validation data are judged manually. The emotion types are the same as comparative studies, divided into seven emotion types, including "angry", "happiness", "thinking", "fear", "sadness", "worry" and "scare". Afterwards, the 500 training articles are imported into the system one by one, so that the system can construct and generate the emotion type coefficient of emotion words automatically. After the inference of emotion word membership coefficient, the emotion types of 1300 test data are deduced one by one. Finally, the inference result and actual result are calculated to obtain the system inference performance, compared with the validation performance of Tung and Lu [4].

Table 5.1: Validation data from "pixnet" blog (2 Actual articles for example)

No.	News Title	News Contents	Actual Emotion Type
1	Back scary jobs	6:00 to get up the cold weather outside Chungli floating drizzle reluctant to leave the warm bed and a warm ~~~~~ no matter how artificial heaters are not willing or want to cheer!	Scare
2	Not saying is only the most sadness	In the end had no idea how it happened, a lot of interpretation is very much like an excuse to accumulate a lot of emotions, the way the article will be afraid of the face, too well, even if it is hope liked this export, because it is really too well	Sadness

(A-2) Comparison with other study - Analysis of validation results

With 500 training articles, the system determines the emotion types of 1300 test data. The system determination result shows in the 1300 test data, there are 1068 actual emotion types deduced correctly, the recall rate, accuracy rate and F-measure are "82.1%". The three indexes deduced by Tung and Lu [4] are "72.5%". Therefore, in the same data and verification process, the inference data of three indicators are slightly higher than Tung and Lu [4] by almost 10% (inference data are shown in Fig. 5.1). However, the validation data, training data and test data used in this paper are different from Tung and Lu [4], all the article data for the validation in this paper are collected "randomly" from blogs, the system inference performance is fair compared with Tung and Lu [4]. According to the aforesaid data, the inference effect of the methodology proposed in this paper and the developed system will be slightly better than Tung and Lu [4].

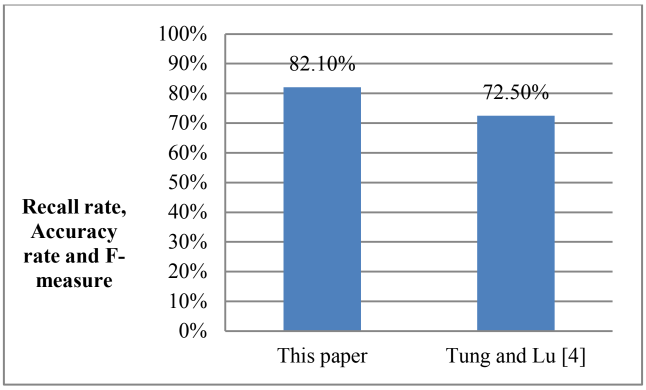


Fig. 5.1: Comparison with Tung and Lu [4]

(B-1) Description of verification procedure of system performance evaluation

To begin with, this paper collects 1000 training articles randomly from "Yahoo News", and 20 Yahoo News articles are selected randomly as test data (e.g. 2 articles in Table 5.2). The test data are the articles voted by over 30 person-times on impressions. In addition, according to the emotions classified by Yahoo News website, there are 7 major emotion types. Afterwards, the system validation process is designed as two stages, in the system validation Stage I, 200 of 1000 training articles are selected randomly and imported into the system, and 20 test data are reasoned, so as to observe the initial determination performance of system. After the Stage I system performance validation is completed, the system test Stage II validation is implemented, this stage is divided into 8 periods, 100 training article data are imported in each period (800 data in all) to analyze the inference result of system in different training test data. The aforesaid 20 test data are used for another inference in each period, so as to analyze the long-term learning trend of system.

Table 5.2: 20 Yahoo News articles (test data, 2 articles for example)

No.	News Title	News Contents	Actual emotion type
1	In order to receive the 18 percent preferential interest rate	Yangming University Hospital physician Chen Xiudan said today that there is an old principal 7-8 years by the respirator…	Sadness
2	"Grandma Sling" affectionate no regrets	GUO Xiu 64-year-old woman suffering from multiple discretion to take care of disabled 12-year-old granddaughter…	Touching

(B-2) System performance evaluation-Analysis of validation results

The system performance evaluation is divided into "Stage I Stage I validation result analysis" and "Stage II validation result analysis". The system validation process and results of various stages are described below.

Stage I validation result analysis

In Stage I system validation, the system determines 20 actual test articles based on 200 training articles, the emotion type inference average recall rate, accuracy rate and F-Measure are 45%, there are 20 actual article emotion types, the system deduces 20 emotion types, there are 9 deduced correctly. The detailed inference result of this stage and the distribution trend of three indexes are shown in Fig. 5.2. Generally speaking, in this stage, the even distribution trend of recall rate, accuracy rate and F-measure is bipolar. However, more than half of validation data cannot be identified correctly in this stage, so the accuracy rate and performance of article emotion type inference are bad.

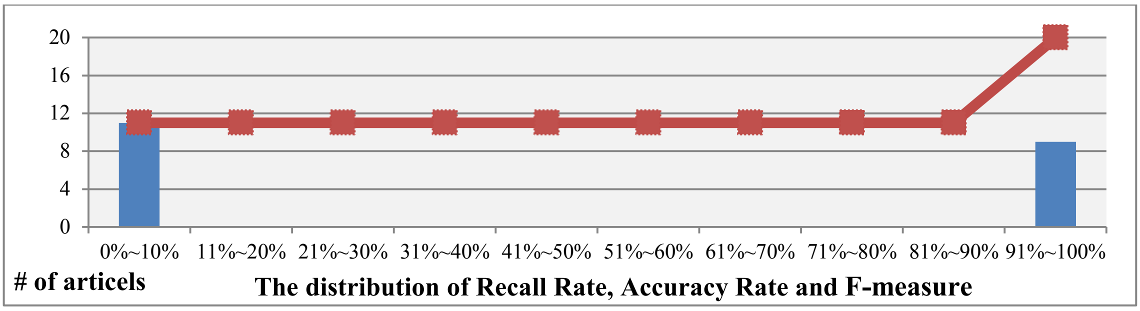


Fig. 5.2: System inference performance of Stage I (The distribution of three indexes)

Stage II: Validation result analysis

Stage II is divided into 8 periods, and 100 training articles are imported into each period, so as to observe the trend (learning behavior) of validation indicator variation in each period as the training articles increase. The validation results of various periods are shown in Table 5.3. According to Table 5.3, the number of training articles is increased by 100 in each period, the overall growth rate of three validation indicators in each period is about 4.4%. In terms of the validation results of the final ninth period (1000 training articles imported), the three indicators are increased from 45% of the first period to 80%. Therefore, the article-expressed emotion determination developed in this paper has learning ability and accuracy.

Table 5.3: The summarization of emotion type inference performance

Emotion type inference: Validation Indicator		**System inference performance of each period- The number of training- articles**									**Avg.**
		Stage I	**Stage II**								
		1st period 200 articles	**2nd period 300 articles**	**3rd period 400 articles**	**4th period 500 articles**	**5th period 600 articles**	**6th period 700 articles**	**7th period 800 articles**	**8th period 900 articles**	**9th period 1000 articles**	
Recall Rate	**Avg.**	45%	50%	55%	60%	65%	70%	75%	80%	80%	64%
	GR.	-	5%	5%	5%	5%	5%	5%	5%	0%	4.4%
Accuracy rate	**Avg.**	45%	50%	55%	60%	65%	70%	75%	80%	80%	64%
	GR.	-	5%	5%	5%	5%	5%	5%	5%	0%	4.4%
F-Measure	**Avg.**	45%	50%	55%	60%	65%	70%	75%	80%	80%	64%
	GR.	-	5%	5%	5%	5%	5%	5%	5%	0%	4.4%
Note: GR refer to Growth Rate											

5.3 Validation of publisher article statement reorganization

The "publisher article statement reorganization module" proposed in this paper reorganizes the statements of violative articles to avoid violating the release specifications. Therefore, the verification samples are constructed before validation, and the verification samples are used as control samples of system validation.

Description of publisher article statement reorganization verification procedure

This paper collects 600 training articles randomly from "Mobile01 forum", in terms of validation data, 20 violative articles are collected randomly from forum (e.g. 2 articles in Table 5.4). The statements of violative articles are segmented by punctuation marks and arranged randomly, to form 20 test articles. Afterwards, the system validation is designed as two stages. In system validation Stage I, 200 of 600 training articles are selected randomly and imported into the system, and 20 test data are used for inference, so as to observe the initial determination performance of system, to observe the accuracy of the methodology proposed in this paper. After the Stage I system performance validation, the system test Stage II is validated, this stage is divided into 4 periods, 100 training articles are imported in each period (600 data in all) to analyze the inference result of system under different training test data. The aforesaid 20 test data are used for another inference in each period to analyze the long-term learning trend of system.

Table 5.4: The violative statements determination questionnaire (No. 1 test article for example)

Sentence No.	Content	Please select a emotion
No. 1	How do you want to fight over China	□ Violative emotion □ Not violative emotion
No. 2	But the military said in a media inquiry	□ Violative emotion □ Not violative emotion

The "actual statement number to be rewritten" of test data should be constructed before validation as the control sample of system inference. Therefore, 30 students who have used information forums for long and have information background, familiar with forum release specifications (from Department of Computer Science and Information Engineering and Department of Information Management) are invited as initial subjects. The statements which may be classified as violative articles are selected from 20 test articles, so as to form the actual violative statements of test articles. To validate the subjective difference and consistency of subjects in reading violative article statements, this paper refers to and improves the online questionnaire designed by Huang [8], based on all the statements in 20 violative articles, to form the violative statement judgment questionnaire for various test articles (e.g. 2 samples in Table 5.4), one repetitive test article is combined randomly (statement number is average of statement number of 20 test articles), the repetitive test article is arranged in 20 test articles randomly, and the subjects evaluate the "actual violative statements" for the 20 test articles, referring to Lee [11], the Root Mean Square (RMS) is used to test the subjective difference of subjects, the "subject repeatability", "subject accuracy" and "overall subjective difference value" are used to evaluate the "subjective consistency and difference" of 30 initial subjects in reading statements. Finally, the first 20 subjects of "overall subjective difference value" are used as final subjects (test results are shown in Table 5.5, case study of 10 subjects), and the violative statements selected by these subjects are used as actual violative statements of 20 validation articles. The three subjective evaluation indexes of subjects are defined and described below:

The number of tests is OTS, the subject repeatability index ORE_i (Equation (5.1)) is the difference average of "violative statement number ORF_i selected by subjects for the first time" and "coincident number ORS_i of violative statements selected by subjects for the second time and the violative statements selected for the first time"; the subject accuracy index OAC_i (Equation (5.2)) is the difference average of "violative statement number OAF_i selected by subjects for the second time" and "coincident number OAS_i of violative statements selected by subjects for the first time"; the overall subjective difference index OSU_i of subjects (Equation (5.3)) is the ratio of "subject repeatability multiplied by subject accuracy and multiplied by 2" to "subject repeatability plus subject accuracy". This index is expected to evaluate the subjectivity of subjects, the lower subjectivity (i.e. closer to 0) represents good subjective feeling of subjects. The symbols are defined as follows:

$$ORE_i = \sqrt{\frac{(ORF_i - ORS_i)^2}{OTS}} \quad (5.1)$$

$$OAC_i = \sqrt{\frac{(OAF_i - OAS_i)^2}{OTS}} \quad (5.2)$$

$$OSU_i = \frac{2 \times ORE_i \times OAC_i}{ORE_i + OAC_i} \quad (5.3)$$

Table 5.5: The subjective difference test result of 10 subjects

Subjects	1	2	3	4	5	6	7	8	9	10
ORF$_i$	64	36	41	54	60	50	52	57	63	61
OAF$_i$	47	58	54	61	58	62	54	61	60	59
ORS$_i$	30	30	37	45	50	39	47	39	56	48
ORE$_i$	24.0	4.2	2.8	6.4	7.1	7.8	3.5	12.7	4.9	9.2
OAC$_i$	12.0	19.8	12.0	11.3	5.7	16.3	4.9	15.6	2.8	7.8
OSU$_i$	16.03	6.99	4.58	8.15	6.29	10.52	4.12	14.00	3.60	8.43
Ranks	2	13	19	12	16	7	21	3	24	11
Not adoption	**Not adoption**					**Not adoption**		**Not adoption**		

Analysis of validation results

The system performance evaluation is divided into "Stage I Stage I validation result analysis" and "Stage II validation result analysis". The system validation process and results of various stages are described below.

Stage I validation result analysis

In Stage I system validation, based on 200 training articles, the system implements inference from 20 actual test articles, the rewritten statement inference average recall rate is 46.61%, accuracy rate is 51.08% and F-measure is 47.22%. Generally speaking, in this stage, the even distribution trend of recall rate, accuracy rate and F-measure is mostly below 50%. However, more than half of validation data cannot be identified correctly at present. Therefore, according to the three indexes obtained from the validation results of this stage, the accuracy rate and performance of publisher article statement reorganization inference are bad.

Stage II validation result analysis

Stage II is divided into 4 periods in this paper, 100 training articles are imported in each period to observe the trend of validation indicator variation in various periods as the training articles increase. The validation results of various periods are shown in Table 5.6.

According to Table 5.6, 100 training articles are increased in each period, the overall growth rate of recall rate in each period is about 7.98%, the accuracy rate is 6.88%, F-measure is 7.58%. In terms of the validation results of the fifth period (600 training articles imported), the recall rate is increased from 46.61% of the first period to 78.51%, the Accuracy rate is increased from 51.08% to 78.58%, the F-measure is increased from 47.22% to 77.53%. Therefore, the publisher article statement reorganization inference developed in this paper has learning ability and considerable accuracy.

Table 5.6: The summarization of Publisher article statement reorganization performance

Publisher article statement reorganization-Validation Indicator		System inference performance of each period- The number of training articles					Avg.
		Stage I	Stage II				
		1st period 200 articles	2nd period 300 articles	3rd period 400 articles	4th period 500 articles	5th period 600 articles	
Recall Rate	Avg.	46.61%	58.28%	66.22%	77.89%	78.51%	65.50%
	GR.	-	11.67%	7.94%	11.67%	0.62%	7.98%
Accuracy rate	Avg.	51.08%	60.63%	68.88%	78.49%	78.58%	67.53%
	GR.	-	9.55%	8.25%	9.61%	0.09%	6.88%
F-Measure	Avg.	47.22%	57.88%	65.97%	77.15%	77.53%	65.15%
	GR.	-	10.66%	8.09%	11.18%	0.38%	7.58%

6.0 CONCLUSION

In recent years, with rapid development of the Internet and rise of forums, the users can publish various articles and discuss the content easily according to speaking specifications and the approval of forum administrators. However, the article writers' lengths and words are different, in order to keep the quality of forum articles, the forums use the existing text matching technology and release specifications to check the article content to filter violative articles. However, the Internet creates so many strange words, when the articles have particular wording, the effect of text matching method of community on filtering violative articles is reduced. Therefore, the forum administrators need to review all articles one by one to filter or revise violative articles, but this action takes a great deal of manpower and time. The article publishers (article providers) may write extreme words in articles unconsciously due to personal carelessness, violating the rules that would be removed by the forum platform or forum administrators, reducing the willingness of article publishers to release articles again.

Therefore, this paper develops an "Article Reorganization Model based on Emotion implied in Forum Articles" to solve the problems in the existing violative article verification mechanism of forums, and further to revise the violative article content immediately. It is based on basic information of forum articles, with statement similarity analysis technique, automatic emotion membership coefficient creation technology, statement review technology, creation of multiple combined sentences, the emotion information implied in articles can be obtained and the content of article statements can be reorganized according to the statements with extreme emotions to assist the forum users and administrators to avoid the articles violating the forum specifications. On the other hand, besides the methodology proposed in this paper, this paper develops a Web-based "article reorganization system for case validation to confirm the feasibility of methodology and technology.

REFERENCES

[1] B. Guido, and V. D. T. Leendert, “Security Policies for Sharing Knowledge in Virtual Communities”. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, Vol. 36 No. 3, 2006, pp. 439-450.

[2] C. H. Liao, B. C. Kuo, and K. C. Pai, “Effectiveness of Automated Chinese Sentence Scoring with Latent Semantic Analysis”. *The Turkish Online Journal of Educational Technology*, Vol. 11, No, 2, 2012, pp. 80-87.

[3] C. L. Hsu, and C. C. Lin, “Acceptance of Blog Usage: The Roles of Technology Acceptance, Social Influence and Knowledge Sharing Motivation”. *Information & Management*, Vol. 45 No. 1, 2007, pp. 65-74.

[4] C. M. Tung, and W. H Lu, “Predict Depression Tendency of Web Posts Using Negative Emotion Evaluation Model”. *In ACM SIGKDD Workshop on Health Informatics*, 2012.

[5] C. Y. Wang, H. Y. Yang, and S. C. T. Chou, “Using Peer-To-Peer Technology for Knowledge Sharing in Communities of Practices”. *Decision Support Systems*, Vol. 45 No. 3, 2008, pp. 528-540.

[6] D. R. David, M. Fernando, L. H. Ramon, M. Stephan, G. Ricardo, M. Cerferino, and D. P. Francisco,

"Conflict And Cognitive Control During Sentence Comprehension Recruitment of a Frontal Network During the Processing of Spanish Object-First Sentences". *Neuropsychologia*, Vol. 49 No. 3, 2011, pp. 382-391.

[7] G. Li, and F. Liu, "Application of a Clustering Method on Sentiment Analysis". *Journal of Information Science*, Vol. 38 No. 2, 2012, pp. 127-139.

[8] H. H. Huang, "Text Analysis of the Relationship between Emotional Text on Facebook Wall". *Department of Information and Learning Technology, National University of Tainan, Master Thesis*, 2013.

[9] L. F. Lai, "A Knowledge Engineering Approach to Knowledge Management". *Information Sciences*, Vol. 177. No. 19, 2007, pp. 4072-4094.

[10] M. Alavi, and D. E. Leidner, "Review Knowledge Management and Knowledge Management Systems Conceptual Foundations and Research Issues". *MIS Quarterly*, Vol. 25 No. 1, 2012, pp. 107-136.

[11] M. F. Lee, "The Influence of Texture-The Influence of Texture-Combinations upon Affective Feelings". *Department of Industrial Design, Tatung University, Master Thesis*, 2010.

[12] P. S. Gregory, and N. A. Daniel, "Emotion Intensity and Categorization Ratings for Emotional and Nonemotional Words". *Cognition and Emotion*, Vol. 22 No. 1, 2007, pp. 114-133.

[13] S. L. Ge, and X. X. Chen, "Cluster Analysis of College English Writing in Automated Essay Scoring". *Computer Engineering and Applications*, Vol. 45 No. 6, 2009, pp. 145-148.

[14] X. Y. Hao, J. H. Li, L. P. You, and K. Y. Liu, "A Research on Building of Chinese Reading Comprehension Corpus". *Journal of Chinese Information Processing*, Vol. 21 No. 6, 2007, pp. 29-35.

[15] Y. Gu, and R. L. Grossman, "Sector: A High Performance Wide Area Forum Data Storage and Sharing System". *Future Generation Computer Systems*, Vol. 26 No. 5, 2010, pp. 720-728.

[16] Y. H. Fang, and C. M. Chiu, "In Justice We Trust: Exploring Knowledge Sharing Continuance Intentions in Virtual Communities of Practice". *Computers in Human Behavior*, Vol. 26 No. 2, 2010, pp. 235-246.

[17] Y. H. Kuo, and H. H. Huang, "Automatic Extraction of Key Sentences via Word Sense Identification for Chinese Text Summarization". *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 11 No. 4, 2007, pp. 416-422.

[18] Y. Ichifuji, S. Konno, and H. Sone, "An Advisory Method for BBS Users and Evaluation of BBS Comments". *Procedia Social and Behavioral Sciences*, Vol. 2 No. 1, 2010, pp. 218-224.

[19] Y. Ko, and J. Seo, "An Effective Sentence-Extraction Technique Using Contextual Information and Statistical Approaches for Text Summarization". *Pattern Recognition Letters*, Vol. 29 No. 9, 2008, pp. 1366-1371.

[20] Y. Lin, Q. Ye, J. Li, Z. Zhang, and T. Wang, "Snippet-Based Unsupervised Approach for Sentiment Classification of Chinese Online Reviews". *International Journal of Information Technology & Decision Making*, Vol. 10 No. 6, 2011, pp. 1097-1110.

[21] Y. M. Li, T. F. Liao, and C. Y. Lai, "A Social Recommender Mechanism for Improving Knowledge Sharing in Online Forums". *Information Processing and Management*, Vol. 48 No. 5, 2012, pp. 978-994.

[22] Y. Miao, J. Su, S. Liu, and J. Zhang, "Bootstrapping-Based Method for Chinese Sentiment Lexicon Construction". *International Conference on Information Engineering Lecture Notes in Information Technology*, Vol. 25, 2012, pp. 248-253.

[23] Y. Y. Chen, C. L. Liu, T. H. Chang, and C. H. Lee, "An Unsupervised Automated Essay Scoring System". *Intelligent Systems*, Vol. 25 No. 5, 2010, pp. 61-67.

[24] Y. Y. Zhao, Q. Bing, and L. Ting, "Integrating Intra- and Inter-Document Evidences for Improving Sentence

Sentiment Classification". *Acta Automatica Sinica*, Vol. 36 No. 10, 2010, pp. 1417-1425.

[25] Z. Halim, and S. Khan, "A Data Science-Based Framework to Categorize Academic Journals". *Scientometrics*, Vol. 119, 2019, pp. 393–423.

[26] B. Rehmam, Z. Halim, G. Abbas, and T. Muhammad, "Artificial Neural Network-Based Speech Recognition Using DWT Analysis Applied on Isolated Words from Oriental Languages". *Malaysian Journal of Computer Science*, Vol. 28 No. 3, 2015, pp. 242-262.

[27] B. Rehman, Z. Halim, and M. Ahmad, ASCII Based GUI System for Arabic Scripted Languages: A Case Study of Urdu". *The International Arab Journal of Information Technology*, Vol. 11 No. 4, 2014, pp. 329-337.