DISTANCE-LEARNING COURSEWARE DISCRIMINATION FOR DISCERNMENT ABILITY TRAINING

Changjie Tang, *Sichuan University, China*

Rynson W.H. Lau, *City University of Hong Kong, Hong Kong*

Qing Li, *City University of Hong Kong, Hong Kong*

Jean W.H. Poon, *City University of Hong Kong, Hong Kong*

Tianqing Zhang, *Sichuan University, China*

Abstract: Training students the discernment ability is an important task in distance education. To improve the training results, this paper makes following contributions: (1) suggests the web courseware on controversial social issues with different viewpoints for discernment training. (2) To construct training courseware, a method called DRPA (Discriminating via Representative Phrase Assaying) is presented with five algorithms, i.e. algorithms for extracting representative phrases, calculating characteristic array, determining the threshold array, objective judgment, and subjective judgment. (3) A new concept called Gymnastics Threshold is proposed and proved to be more accuracy than traditional threshold. (4) Extensive experiments are given to show that: DRPA is much more efficient and accuracy than traditional method.

Keywords: Distance learning, discernment ability training, Automatic document discrimination, natural language understanding.

BACKGROUND AND PROJECT MOTIVATION

As an experiment of the distance learning M.Sc. and Ph.D. degree programs, one of the authors is currently teaching, partially via the Internet, a course named “Reading Selected Articles on the Web” (RSAW) to students across several provinces in China. In order to improve the learning quality and the performance of tutoring system, we have developed the Web Tutor Object Tree (WTO Tree) (Tang C. Lau R.W.H. Li Q.,Yin H. Li T. and D. Kilis. 2000)., a method to construct personalized courseware to adapt the teaching according to the ability of individual students. The popularity of distance learning has prompted the demands on web courseware discrimination, as illustrated in the following examples:

Example 1. The International Debate Competition for University Students (IDCU) is being held annually in Asia and the Pacific region as well as other areas of the world. The competing teams will draw cuts to determine their points of view (Pro or Con). After drawing cuts, the teams have 24 hours to prepare their debate. In order to use the web resources, the coaches and teams are in urgent need of a tool to help discriminate web courseware with different points of view on a specific topic.

Example 2. In a research area, there are often different academic camps holding different or opposed view points, for example, the arguments on human rights in Africa, the debate on heredity and aberrance in biology, the arguments about “goto” statement in programming languages, the controversy about universal schema in database area, etc.. In order to train the discernment ability of distance learning students, professors teaching the RSAW course need a tool to recognize and select web documents in different view points to organize the Web Tutor Tree (Tang C. Lau R.W.H. Li Q.,Yin H. Li T. and D. Kilis.2000).. Sometimes, a professor may prefer his/her students to read 60% read 60% of the papers in a particular point of view and 40% of papers in the other point of view. •

The above examples show that the training student’s discernment ability is an important task for distance education. To enhance student’s discernment ability by distance education, we suggest the web courseware on controversial social issues with different viewpoints for discernment training. In order to construct good training courseware for this purpose and to check students’ exercises in discernment training, an automatic document discrimination method called DRPA (Discriminating via Representative Phrase) is proposed along with five algorithms in this paper.

The rest of the paper is organized as follows. Section 2 gives related works and deficiency of existing work, Section 3 discusses the limitation of traditional approaches for courseware discrimination and discusses the special approach in this work. Section 4 gives preliminary concepts and notions of DRPA. Sections 5 and 6 present the algorithms for extracting representative phrases, calculating characteristic array and threshold array, objective discrimination, and subjective discrimination. Section 7 shows some experimental results of the proposed method in a classroom environment, demonstrating the efficiency of the algorithms. Finally, section 8 briefly concludes the paper.

RELATED WORK AND DEFICIENCY OF TRADITIONAL METHODS

Document discrimination can be viewed as a special case of text classification that is widely used in e-commerce, e-services, virtual offices and network security (e.g., discriminating black emails), and has been studied extensively as many methods have been developed [Tang C., Lau R.W.H., Yin H. Li Q. Lu Y. Yu Z., Xiang L., and Zhang T. 1999), Tang C., Yu Z., You Z., Zhang T., and Yang. L. 2000, Mannila H. and Toivonen H.1999), M. Jiang, S. Tseng, and C. Tsai.1999). Various research directions have been explored, such as K-nearest neighbor (kNN) algorithms, neural networks , decision trees (ID3)(Damerau C., Weiss F.1994) , rule learning, support vector machine (SVM), linear classifiers, and Naive Bayes methods. Some new methods have been studied in recent years for text classification (Tang C, Li Q. Rynson W.H.L. Huang X .2003). As the limit of paper space , the detail analyses and comparison for the related works are omitted here. By the special feature, the general text classification methods are not efficient for Web documents discrimination. Most existing methods discriminate documents based on keyword matching. Some of them may require human interaction. This raises two problems:

（1）Low efficiency caused by the interactive process and the manual operation.

（2）Low accuracy of identification including failure to report con (negative) courseware and treating pro (positive) as con courseware.

A Web courseware is a set of semi-structural files (usually in the format of HTML, XML, PDF or emails) containing formatted elements and non-formatted elements, HyperText Links. The formatted elements include header, title, URL, author, and the other extractable parts by XML technique. The rest, i.e., non-formatted elements, are as heap of sentences. The Semi-structure feature has been studied extensively. The traditional keyword-matching approach includes two steps:

a. Handling formatted elements

The formatted part, such as header, title, author, URL, and email address, is stored as a set of database records in the format of (<header, weight_h>, <name, weight_n>, <author, weight_a>, <URL, weight_u>, ...), where weight_h is the weight indicating the contribution of the header to the classification, and so on. The weight values are initialized by the expert’s experience and re-evaluated after each session, i.e., they are being updated in the learning process. The process of handling formatted part is relatively mature, and hence will not be discussed in this paper.

b. Handling non-formatted elements

Since the non-formatted part is a set of sentences, the traditional processing methods to discriminate courseware usually include the following steps:

- Predefine a set of sensitive words.
- Extract sensitive words from the documents and compute the number of occurrences. If this number of occurrences is greater than a predefined threshold, the courseware is classified as Con (point of view);

To overcome these limitations of traditional key words matching, this paper adopts the technique from natural language understanding. The details will be explained in rest sections.

SPECIAL APPROACH FOR WEB COURSEWARE DISCRIMINATION

The discrimination of Web courseware is different from general document discrimination because Web courseware can be well organized into a Web Tutor Tree(Tang C. Lau R.W.H. Li Q.,Yin H. Li T. and D. Kilis. 2000) under a predefined topic. In addition, the different viewpoints are known and may be expressed by representative (pro/con) sentences. To overcome the special difficulties in Web courseware discrimination described above, this paper proposes a new method named *DRPA* (Discriminating web courseware based on Representative Phrase Assaying). First, DRPA extracts the representative sentences (such as the first and last sentences in a section) from a training set of web courseware. Second, it parses these sentences to get the main semantic elements. Third, it counts and assays pro/con elements. Finally, by employing a typical data mining method, DRPA gets the judgment rules for Web courseware.

CONCEPTS AND NOTIONS

Table 1 provides a summary of the notations and terminologies used in this article. The concepts of the representative phrase and gymnastics average threshold are the special contributions of this work.

Symbol	**Explanation**
TrainWareS	A set of Web courseware selected by the user satisfying the ENOUGH Criteria; see definition 1.
TrainPhrSet	The set of representative phrases of all the document in TrainWareS
Sub, Pred and Obj.	Subject, predicate and object in a sentence of natural language
RepSentenSet	The set containing the first and last sentences of each paragraph, and sentences with emphatic marks.
RepSenten	Member of RepSentenSet
GetRepSentenSet(*d*, *r*)	Procedure to get a RepSentenSet from document ***d***.
Con	Con ratio of *d*; see definition 4
Pro	Pro ratio of *d*; see definition 4
PhraseWeight	A number indicating the contribution of the sentence to the classification and is initialized at the training stage.
Score	The average score of all phrases of ***d***.
CA	The characteristic array of document ***d***, CA = <Score, Con, Pro>
GymnasticsAve(*S*)	The gymnastics average value of S, See definition 5
Opt_threshold	Optimistic threshold array of characteristics array featured by minimum Con and Pro; see definition 6
Pes_threshold	Pessimistic threshold array of characteristics array featured by maximum Con and Pro; see definition 6
Gym_threshold	Gymnastics threshold array of characteristics array featured by gymnastics average; see definition 6

Table 1. The notions and terminology used in this paper.

The major deficiency of the traditional approach is that it only considers isolated keywords during the matching process, instead of the meaning of the Web courseware. Hence, it is a lexical but not semantic mechanism. This paper takes the semantic approach. To formalize it, we give the following definitions.

Definition 1. Let TrainWareS be a set of Web courseware selected by the user satisfying the following criteria (called Enough Criteria):

1. TrainWareS should be big enough for the discriminating task, say containing more than 200 Courseware documents.
2. Each document in TrainWareS is already classified (pro or con).
3. It should be representative enough for the task, say, containing the typical pro and con classes of courseware documents.

Experiments conducted on natural language processing show that some sentences condense the idea of a paragraph. To formalize this observation, we define the following concepts:

Definition 2. The representative sentence set of a document, denoted as RepSentenSet, is the set containing the first and last sentences of each paragraph, all sentences from the first and the last paragraphs in that document, and sentences with emphatic marks. An element of RepSentenSet is denoted as RepSenten.

The procedure GetRepSentenSet(*doc*, RepSentenSet) is to obtain the set of representative sentences from a courseware document *doc*; this is a simple procedure, which details are omitted. In Web files, a paragraph is indicated by HTML or XML marks such as <p>, </p>, <br>, etc..

Linguistics research shows that 70% or more meaning of a sentence can be expressed by SPO-Structure (i.e., the expression of Subject, Predicate and Object). Thus, we have:

Definition 3. Let *S* be an element in RepSentenSet. Its subject, predicate and object are denoted as Sub, Pred and Obj, respectively. Let Φ be the empty word. Then:

1. The set of 4-tuples {(Sub, Pred, Φ, PhraseWeight), (Φ, Pred, Obj, PhraseWeight), (Sub, Pred, Obj, PhraseWeight)}, denoted as RepPhrase, is called the ***representative phrase*** of *S*, where PhraseWeight is a number indicating its contribution to the classification and is initialized at the training stage.
2. The representative phrases of document doc are denoted as TrainPhr(*doc*). The set of representative phrases of all the document of TrainWareS, denoted as TrainPhrSet, is called the ***training phrase set.***

Example 3. Consider the Web courseware for distance learning students in Economics. It includes documents about stocks from *Yahoo!Finance* (Zuo J, Tang C. and Zhang T.2002). Representative sentences of document doc are listed below:

S1: The markets were up after a government report showed job growth and wages rose less than expected in May, raising optimism that the Federal Reserve can limit the number of interest-rate increases in coming months.

S2: After some softness early in the morning, the Dow has scaled back from a loss while the Nasdaq and S&P added to their gains.

S3: The Nasdaq Composite rose 85.17 points to 4098.51 and it seems that the market can maintain the gains.

S4: The S&P 500 rose 10.67 points to 1502.92.

The corresponding representative phrases are shown in Table 2.

Sentence	**Subject**	**Predicate**	**Object**
1	Markets	Were	Up
2	Dow	Scaled	Back
3	Nasdaq	Rose	
4	S&P 500	Rose	

Table 2. Samples of representative phrases

Intuitively, a courseware document is evaluated synthetically by assaying all phrases such as the number of Pro / Con phrases, and the average score of the phrases. To formalize this, we need to define the characteristic array of a document.

Definition 4 (Characteristic array of a document) Let *doc* be a courseware document, and $\{p_1, p_2, ..., p_n\}$ be the set of representative phrases extracted from *doc*, with *n* being the size of the set. Let Negative and Positive be the numbers of all Con phrases and Pro phrases, respectively.

1. The con ratio of *doc* is defined as: Con = -Negative / *n*.
2. The pro ratio of *doc* is defined as: Pro = Positive / *n*.
3. The average score of all phrases of *doc* is defined as: $Score = \sum_{i=1}^{n} p_i * ContriPhr / n$.
4. The characteristic array of document doc is defined as: CA = <Score, Con, Pro>.

Note that Negative ≥ 0, Con ≤ 0, and the score in the characteristic array of *doc* is different from that of the representative phrase denoted by PhraseWeight.

Definition 5 (Gymnastics Average) Let *S* be a set of real numbers. The gymnastics average value of *S*, denoted as GymnasticsAve(*S*), is the average value of the elements in S_1, where S_1 is the result of deleting 10% elements at the beginning and 10% elements at the end of *S*.

Note that the gymnastics average value is just like the scoring rule in such sport games as gymnastics and diving.

Definition 6 (Threshold Array) Let TrainWareS be the training set, CA_Set = {CA_i | CA_i is a characteristic array of document d_i, $1 \leq i \leq n$ } be the set of characteristic arrays of TrainWareS.

1. The optimistic threshold array of CA_Set is defined as Opt_threshold(CA_Set) = (Con_threshold,

Pro_threshold, Score_threshold), where Con_threshold = Min{CA$_i$.Con | CA$_i$ in CA_Set}, Pro_threshold = Min{CA$_i$.Pro | CA$_i$ in CA_Set}, and Score_threshold = Average({CA$_i$.Score | CA$_i$ in CA_Set}).

2. The pessimistic threshold array of CA_Set is defined as Pes_threshold(CA_Set) = (Con_threshold, Pro_threshold, Score_threshold), where Con_threshold = Max{CA$_i$.Con | CA$_i$ in CA_Set}, Pro_threshold = Max{CA$_i$.Pro | CA$_i$ in CA_Set}, and Score_threshold = Average({CA$_i$.Score | CA$_i$ in CA_Set}).
3. The Gymnastics threshold array of CA_Set is defined as Gym_threshold(CA_Set) = (Con_threshold, Pro_threshold, Score_threshold), where Con_threshold = GymnasticsAve({CA$_i$.Con | CA$_i$ in CA_Set}), Pro_threshold = GymnasticsAve({CA$_i$.Pro | CA$_i$ in CA_Set}), and Score_threshold = GymnasticsAve({CA$_i$.Score | CA$_i$ in CA_Set}).

Observation 1. In the web leaning environment, let TrainWareS be the training set, CA_Set={CA$_i$ | CA$_i$ is a characteristic array of document d$_i$, $1 \le i \le n$ }, GymAveCon = GymnasticsAve({CA$_i$.Con| $1 \le i \le n$ }), and GymAvePos = GymnasticsAve({CA$_i$.Pos| $1 \le i \le n$ }). From the experimental statistics, it is true in most cases that:

(a) Let i=Min{k | CA$_k$.Con $\le$ GymAveCon $\le$ CA$_{k+1}$.Con}. Then k is in the middle 10% of interval [0, n].

(b) Let j=Min{k | CA$_k$.Pos $\le$ GymAvePos $\le$ CA$_{k+1}$.Pos}. Then k is in the middle 10% of interval [0, n].

(c) Max{CA$_i$.Con | CA$_i$ in CA_Set} μ 110% * GymAveCon.

(d) GymAveCon μ 90% * Min{CA$_i$.Con | CA$_i$ in CA_Set}.

(e) Max{CA$_i$.Pro | CA$_i$ in CA_Set} μ 110% * GymAvePos.

(f) GymAvePos μ 90% * Min{CA$_i$.Pro | CA$_i$ in CA_Set}.

This observation is based on statistical results and can be explained as follows. The teachers select web documents as course material according to the natural distribution of the score (Pro and Con).

Proposition 1. Suppose that the above observation is true for TrainWareS and n $\ge$ 10. Then the Gymnastics threshold is more accuracy than the pessimistic threshold and the optimistic threshold.

Proof. The general proof for arbitrary n is prolixity and fussy. Here, we give a proof sketch for the statement on optimistic threshold, while the number of documents n=10. The principle for other cases is expected to be the same. It is easy to see that adding a very small positive number (close to zero), έ, to CA$_i$.Con does not affect the whole statistical property. Hence, we may assume that no two optimistic threshold numbers are equal. Without loss of generality, we assume CA$_1$.Con > CA$_2$.Con > ... > CA$_{10}$.Con. (This assumption simplifies the proof.) For simplicity, we denote

C$_i$ = CA$_i$.Con,

Optimistic Con_threshold as OptCon = Min$\{CA_i.Con \mid CA_i$ in CA_Set, $1 \le i \le 10\} = C_{10}$,

g = GymnasticsAve $\{CA_i.Con \mid CA_i$ in CA_Set, $1 \le i \le 10\}$,

OptimiticsDisSum = $|CA_2.Con - OptCon| + |CA_3.Con - OptCon| + \dots + |CA_9.Con - OptCon|$.

Note here that the top 10% of $CA_a.Con$ is ignored.

GymnasticDisSum = $|CA_1.Con - GymCon| + |CA_2.Con - GymCon| + \dots + |CA_9.Con - GymCon|$

Then $g = (C_2 + C_3 + \dots + C_9) / 8$. It is easy to see that $C_2 \ge g \ge C_9$.

Suppose $C_i \ge GymCon \ge C_{i+1}$, where $1<i<10$.

Let Delta = $(10-2*i) * GymCon + 8*OptCon + 2*(C_{i+1} + \dots + C_9)$, then

OptimiticsDisSum – DistanceGymSum $= (C_2 – C_{10}) + (C_3 - C_{10}) + \dots + (C_i - C_{10}) + (C_{i+1} - C_{10})$

$+ \dots + (C_9 - C_{10})$

$- [(C_2 - g) + (C_3 - g) + \dots + (C_i - g)] - [(g - C_{i+1}) + \dots + (g – C_9)]$

$= (10-2*i) * GymCon + 8*OptCon + 2*(C_{i+1} + \dots + C_9)$

= Delta

By (a) in Observation 1, i is around 4 or 5 (10% in the interval of [0,10]). Thus Delta=8*OptCon + $2*(C_{i+1} + \dots + C_9) > 0$. This is true in most cases for the Web leaning environment under lemma's assumption. Similar proof can be applied to optimistic Pro, pessimistic Con and pessimistic Pro. Thus the Gymnastics threshold is more accurate than the pessimistic threshold and the optimistic threshold, as desired.

The assumption for $n \ge 10$ is used to simplify the statement concerning the Gymnastic Average procedure. The case $n<10$ is meaningless in practice.

Based on the above concepts, the main idea of our assaying mechanism can be described by the following two steps:

1. *Learning*: Get the characteristic array from TrainWareS
    - Construct TrainWareS of courseware documents according to the ENOUGH Criteria.
    - Distill TrainPhr from TrainWareS by parsing technique.
    - Evaluate each phrase in TrainPhr by setting p.PhraseWeight to the default (expert defined) value.
    - Calculate the characteristic array for each courseware document in TrainWareS.
    - Mine the threshold array from the set of characteristic arrays of TrainWareS.
2. *Application*: Classify the specified courseware document *doc*
    - Extract representative phrases from document *doc* by parsing technique.
    - Get p.PhraseWeight by matching p in TrainPhrSet.
    - Calculate characteristic array CA of *doc.*
    - Classify *doc* based on comparing CA with the threshold array of TrainWareS.

DISTILLING REPRESENTATIVE PHRASES

The grammatical analysis algorithm for representative sentences is borrowed from the *XinYi Internet Chinese-English Translation System.* It is a software product developed by us (Zuo J, Tang C. and Zhang T. 2002). The syntax knowledge, including sentence structure rules and the conjunction rules of verbs, is implemented as knowledge bases *KB1* and *KB2*, respectively, in the XinYi system. The system uses verbs as the kernel of each sentence. It can analyze the semantics of verb tense. It can extract phrases such as (Subject, Predicate) and (Predicate, Object). The classification rules borrowed from the Oxford Dictionary form the main part of the verb rule base. The verbs are divided into 25 classes, which can be expressed as VP1, VP2, …, VP25. More details on this can be found in the prefix of the Oxford dictionary.

Example 4. Consider the syntax rules in XinYi system: S(VERB1) → SUB + VERB1 + NOUN. This is a representative rule in KB1. It describes the sentence structure with a verb. It requires the structure to be something like "Subject be Object". The sentence "It is a big room." satisfies this rule.

Example 5. Consider the rule: PREPNO → PREP + NOUN. This is a representative rule in KB2. It describes the phrase structure in the form of "Preposition + object". The object is a word or a Noun phrase. Based on the syntax knowledge bases KB1 and KB2, the bottom-up parsing methods are used to distill the grammatical elements such as subject, predicate and object from the sentence. In this way, we get the representative phrases. Algorithm 1 shows the process to distill representative phrases from a document.

Algorithm 1 (Distill representative phrase)
Input: Courseware document *doc*.
Output: Representative phrase set RepPhraseSet of *doc*.
Steps:

```
    GetRepSentenSet(doc, RepSentenSet);
    RepPhrasesSet = Empty;
    for each RepSenten in RepSentenSet {
        Sentence_Pattern = Get_Pattern_From_KB1(Verb);  // Verb is kernal of the pattern.
        // the result is: Sentence_Pattern = Subject_Phrase + Verb + Object_Phrase
        Subject = Extract_Subject(Subject_Phrase);       // such as get Noue "Adjective+None"
        Oubject = Extract_Oubject(Object_Phrase);        // such as get Noue from: "Adjective+ None+Adverb"
        RepPhrase = (Subject,Verb,Object);               // Subject or Object may be $\Phi$;
        RepPhraseSet = RepPhraseSet + RepPhrase;
    }
    Output RepPhraseSet;
```

Algorithm 1 is non-determinate. Thus, the back tracing mechanism is used to optimize the parsing process. The max back trace depth is set to 1,000. Our implementation of this algorithm in the XinYi system can process more than 1,500 Chinese characters per minute with CPU working in frequency 400M(Yu Z. Tang C. and Zhang T. 2000).

Observation 2. Let m be the average number of words in sentence of web document. Then m≥2 and 25≥m. This observation is based on the experiments on the real web documents.

Proposition 2. Let n be the number of sentences of input documents. Assume observation 2 is true, then the complexity of Algorithm 1 is O(n).

Proof. By the observation 2, at most n/2 sentences to be analyzed. And at most 25 words in a sentences. Note that in Algorithm 1 the functions GetRepSentenSe is trivial parsing. The other functions such as Get Pattern From KB1, Extract Subject(Subject_Phrase),And Extract_Oubject(Object_Phrase) are working with sentence for at most 25 words. Thus, the cost can be evaluated as 25d for some constant d. The total cost should be O((n/2)*25*d), i.e., O(n).

ASSAYING AND DISCRIMINATION

The characteristic array of a document is in the format CA = <Con, Pro, Score> (see Definition 4). It summarizes the characteristics of a document such as the percentage of Pro/Con phrases in the document. Algorithm 2 describes the process of calculating this characteristic array.

Algorithm 2 (Calculating characteristic array of a document)
Input: Courseware document *doc*, Stage. // Stage is “Training” or “Discriminating”
Output: The characteristic array CA of a document.
Steps:

```
    Get representative phrase set RepPhraseSet from RepSenten by Algorithm 1;
    Initialize characteristic array CA = <Con, Pro, Score> as <0,0,0>;
    TotalScore=0; TotalPharase=0;
    for each phrase p in RepPhrases {          // Get p.Score:
        if (Stage == Training) {
            if (doc is pro) p.PhraseWeight is assigned a positive number by expert experience;
            if (doc is con) p.PhraseWeight is assigned a negative number by expert experience;
        } else if (Stage == Discriminating)
            p.PhraseWeight = TrainPhrSet.p.PhraseWeight;
        TotalScore = TotalScore+p.PhraseWeight;
        TotalPharase = TotalPharase + 1;
        if (p.PhraseWeight > T_Phrase)
            CA.Pro = CA.Pro + 1;
        else  CA.Con = CA.Con -1;
     }
    CA.Score = TotalScore / TotalPhrase;
    output CA;
```

As illustrated in section 3, one of the important steps in the training stage is to get the threshold array from TrainWareS. Algorithm 3 describes this process.

Algorithm 3 (Calculating threshold array)
Input: TrainPhrSet of TainSet, Style. // Style is in {Optimistic, Pessimistic, Gymnastics}
Output: threshold array T_Array of TrainWareS.
Steps:

```
    Calculate the set of characteristic arrays for all documents in TranSet by Algorithm 2. Denote result as VertorSet;
    if  (Style == Optimistic)
        T_Array = Opt_Threshold(VertorSet);
    else if  (Style == Pessimistic)
        T_Array = Pes_Threshold (VertorSet);
    else if  (Style == Gymnastics )
        T_Array = Gymnastics_Threshold (VertorSet);
    Output T_array;
```

In Algorithm 3, the input TrainPhrSet (of the training set) is in the form of TrainPhrSet={<Sub, Pred, Obj, PhraseWeight>}, where PhraseWeight is a value indicating the contribution of the sentence to the classification and is assigned based on expert experience during the data cleaning process. For example, (Dow Jones, goes, up, 16) represents a piece of good news and hence is given a high score. However, (NASDAQ, goes, down, -8) represents a bad news in the stock market.

Now, we present two algorithms to judge the courseware documents. Algorithm 4 shown below is referred to as "objective" because it is based on the objective facts. It only uses the ratio of negative phrases to positive phrases extracted from each document in TrainWareS.

Proposition 3 Let k be number of TrainPhrSet in TainSet. Then the complexity of Algorithm 3 is O(K).

Proof . Note that the statement after "IF" is simple calculation for simple arithmetic calculation. Thus the main cost is in the t step "Calculate the set of characteristic arrays for all documents in TranSet by Algorithm 2 ", By proposition 2.it is linear.Hence the total cost can be evaluated as O(K).

Algorithm 4 (Objective Discrimination algorithm)
Input: The Courseware document *doc* to be judged**,** TrainWareS, Style.
    // Style is in {Optimistic, Pessimistic, Gymnastics}
Ouput: The judging conclusion of *doc*. // "Yes" (*doc* is accepted) or "No" (*doc* is rejected)
Steps:

```
    Extract TrainPhrSet from TainSet by Algorithm 1;
    According to Style, mine the threshold array of TrainWareS by Algorithm 3 and denote as T_Array;
    Extract representative phrases from document doc and denote it as RepPhrases;
    Match each phrase of RepPhrases and get its score;
    if (NOT matched) {
        p.score = 0;
        Get the characteristic array CA of doc by Algorithm 2;
```

```
}
if (T_Array.Pro == 0)     T_Array.Pro = 0.01;       // avoiding zero as the divisor
if (CA.Pro==0)            CA.Pro = 0.01;            // avoiding zero as the divisor
Judge_Ratio = (-CA, Con / CA.Pro);
Threshod_Ratio = (-T_Array.Con / T_Array.Pro)
if (Judge_Ratio >= Threshod_Ratio)
    conclusion = "No";        // reject doc
else  conclusion = "Yes";     // accept doc
```

Algorithm 5 shown below is referred to as “subjective” because it is based on the score reflecting the subjective experience of the expert.

Algorithm 5 (Subjective Discrimination algorithm)

```
Input:  TrainWareS and the courseware document doc to be evaluated, EvaluationStyle;
     // EvaluationStyle is in {Optimistic, Pessimistic, Gymnastics}
Ouput: The Evaluation conclusion of doc;       // Pro or Con
Steps:
    Extarct TrainPhrSet from TainSet by Algorithm 1;
    According to EvaluationStyle, mine threshold array TrainWareS by Algorithm 3. Denote output as T_Array;
    Extract representative phrases from doc and denote as RepPhrases;
    Match each phrase of RepPhrases and get its score;
    if (NOT matched)    p.score = 0;
    Get the characteristic array CA of doc by Algorithm 2;
    if (CA.Score < T_Array.Score)
        conclusion is “Con”;
    else  conclusion is “Pro”;
```

EXPERIMENTAL RESULTS AND ANALYSES

To demonstrate the effect of our algorithms, we have conducted an initial classroom-based experiment. In this section, we report our experimental study in terms of the design and method of analysis.

Objectives of the Classroom Experiment

The experiment was carried out in a classroom because of three reasons:

(1) An in-depth and thorough evaluation method for DRPA requires empirical comparison with human involvement in a real teaching environment, which can reveal the validity and robustness of DRPA.

(2) We want to understand the cognitive process of human beings when they are discerning or grouping the texts. This can help us develop DRPA to perform more “intelligently” and in a more “humanized” manner.

(3) Since the data collected in the classroom is believed to be more realistic and reliable, the result generated can serve as a standard reference for comparison with the one generated by DRPA.

Setup of the Classroom Experiment

Downloaded documents, which were in the area of Economics, were distributed to a group of 30 Economics-majored students. They were asked to read these documents in a one-hour session and to discern the documents into pro and con on this topic; they were asked to underline the sentences that gave hints on their judgment. Their strategies, methods or behavior applied in the reading process were recorded. The experiment was conducted under a controlled environment in an invigilated classroom setting.

The following were the detailed steps of the experiment:

(1) Each person is distributed with a set of 10 pieces of documents.

(2) They are asked to read these articles with their own strategy in an hour.

(3) With the same topic defined in DRPA "The trend of Hong Kong economy", students are asked to discern the documents into pro and con on this topic.

(4) They were asked to underline the sentences that give hints on their judgment.

(5) If the document has positive standpoints, then mark +.

(6) After an hour, the documents are collected.

(7) An interview is conducted after the reading session.

(8) Their strategies, methods or behaviors used in the reading process are recorded.

(9) Their discernment results are determined by analyzing the underlined sentences and the symbolic marks given on each document.

After the experiments were conducted, the importance of each proposition and the performance in their discrimination are evaluated on the basis of how the text was discriminated and marked by the students according to the method discussed in the following section.

Method of Analysis

Results of the classroom experiment are analyzed within the framework in terms of reading frequency, sentence distillation, diagnostic units and the discrimination result.

A. Reading Frequency

To find out the reading frequency before making discernment is an important reference to understand the human reading process.

Findings:

Table 3 shows that students need to read 1.28 times in average before making final discrimination. It tells that, even as intelligent like human beings, we would read a document more than once in order to make a judgment if necessary. In DRPA, the "reading" and "discernment" processes are in one linear flow: distill representative phrase à calculating characteristic array of a document à calculating threshold array à objective discrimination algorithm à subjective discrimination algorithm. Based on the finding about the human reading frequency, we would like to know if it is possible to make the discernment process more sophisticated and result more accurately in the way to “read” the document for over one time or adopt a recursive calculation.

Frequency	1	2	3	4
Percentage	79%	14%	6%	1%

Table 3. The notions and terminology used in this paper

B. Sentence Distillation

After their documents are collected, the underlined sentences which provide hints for discernment were counted and averaged. The result is shown in Table 4.

	DRPA	**Classroom Experiment**
Means	10.2	5.99

Table 4. The average number of distilled sentences

Findings:

From the result shown in Table 4, it is obvious that the number of sentences underlined by students in each document is smaller than that extracted by DRPA. The large number of sentences distilled by DRPA may indicate high degree of redundancy in the process of sentence distillation. The redundant phrases from DRPA may not bring any benefit to increase the hit rate in matching. The low matching rate in DRPA application implies that the locations where the representative sentences distilled suggested in DRPA (paragraph-initiated sentences / paragraph-last sentences / all sentences in first and last paragraphs) may not help get the most relevant and indicative sentences.

C. Diagnostic Units

Findings:

From the result of their marked documents and feedback in the interview, it is found that, within any text, there exists different heuristics that may contain obvious hints and important information helpful for document discrimination. Table 5 is a summary of diagnostic units that the students mostly noticed and used in the reading process:

Table 5. A summary of diagnostic units

Title
Sub-title
Tone of the sentence
Synonyms
First and the last paragraph
Topic sentence
Thematic words

Getting insight from investigating the cognitive process of human being in reading, it is found that the semantic approach suggested in DRPA is not significant enough for corpus analysis. Rather, other components of NLP like the weight of titles, topic sentences in introduction and conclusion are believed as potential heuristics in contributing result with higher accuracy.

D. Discernment Result

Findings:

Table 6 shows that 90% of the documents have the same result in judgment except Econ_C. Since the result generated by the human subjects is more reliable, the result of Econ_C generated by DRPA is believed to be incorrect. Document discernment is the process of spotting relevant hints and information from documents. It involves sophisticated and intelligent manipulation of given knowledge as well as knowledge of natural language (Riloff E. and Lehnert W. 1994). The conclusion we can draw from the results of two experiments is that a system like DRPA, which uses only the existing semantic structure to determine the relative importance of the representative phrases, performs well to a certain extent. However, we have found that redundant and irrelevant phrases distilled in DRPA may not contribute much to correct discrimination. Learned from classroom experiments, we can perform discrimination work more accurately and flexibly by considering a variety of hints or heuristics like the title or the topic sentence. In addition, the experiment tells that we have to extract sentences which are more representative and carries heavier weight than the one suggested in DRPA. This is important as it can minimize the ineffective distillation by extracting too many redundant and "less representative" sentences.

Table 6. The discrimination results of a DRPA and classroom experiment

Document	**DRPA**	**Classroom Experiment**
Econ_A	Pro	Pro

Econ_B	Con	Con	
Econ_C	**Con**	**Pro**	⇐
Econ_D	Pro	Pro	
Econ_E	Pro	Pro	
Econ_F	Con	Con	
Econ_G	Con	Con	
Econ_H	Pro	Pro	
Econ_I	Pro	Pro	
Econ_J	Con	Con	

Discussions

We may now summarize the findings that we have obtained from the experiment:

1. ***Title:*** The title, heading or subheading in a document is a good candidate for discernment. From the definition of linguistics, title is defined as a phrase that summarizes the information of a text in a compressed and condensed way. It conveys the main position and main theme of content. Since the title has the indicative function, it should be extracted and treated as a “representative” phrase in training and discrimination stage.

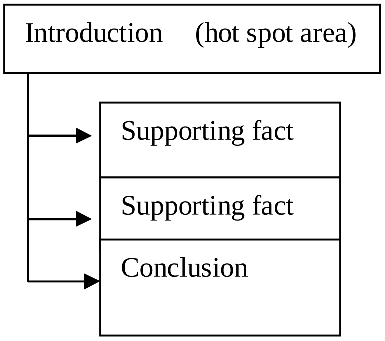
   It is suggested that, in the *training stage*, title and subtitles are extracted and assigned a score for training as usual. This measure can increase the size of the training pool that improves the matching rate in comparison. With the addition of titles, the size of training pool can be increased by 16.3%. This can increase the hit rate in the subsequent comparison.

   In the *discrimination stage*, the title in the testing document is compared and scored for the first priority. If the title has already carried an obvious and indicative score, we may skip the process to extract representative phrases in the body part of document. Finally, based on the score of the title, we calculate its characteristic array and make judgment in the same way as in the original DRPA. Title matching is a reliable and simple method in which it can minimize the redundant representative phrases in the training and discrimination phase. Effective discrimination and accurate result can be achieved.

2. ***Subject Categorization and Pattern Recognition:*** As mentioned above, the dependence on cut off semantic structure may cause it to be too ambiguous to do assaying work. So, apart from semantic analysis, text pattern is another good indicator for classification. In content area, pattern recognition is a tactic for identifying organizational patterns in information we read or hear. It constitutes a key to the

basic understanding of the inherent logical structure underlying, providing potential useful approach for rational classification.

With reference to the principle of content thinking information skills, there is a standardized “relationship” and “matching” between subjects and text pattern. For example, geographic document is always written in the Process Pattern while the historic documents are mostly in Sequence Pattern or Cause and Effect Pattern. After we understand the relationship, we have to identify the “hot spot” in different pattern because most of the relevant and important information are located in specific hotspot according to specific pattern. Figure 1 shows the relationship between the pattern and the location.



For those documents belonging to generalization pattern, the statements in the first paragraph always contain the opinion and the standpoint of an author. Other part like supporting facts which contain data or example are relatively less important for classification.

Conclusion mostly is only the summary of above information of the supporting facts and repeated claiming of theme or standpoint mentioned in the first paragraph. So, the sentences in the first paragraph should be extracted and scored with heavier weigh.

Figure 1. The Generalization pattern related subject: Economics.

Representative phrases, according to the “hot spot method” mentioned above, are extracted in “hot spot” (first paragraph) which should carry a heavy weight for discrimination. If there is no matching or the ratio of zero score in this hot spot is very high, then we extract and match the paragraph-initial and paragraph-final sentences from the rest of the document. Based on this summary, we may refine our method with the title method and “hot spot” approach as follows:

Refined Step:

```
if (title not matched) {
    Extract representative sentence in hot spot;
    If (matching found)
        Scoring and Discrimination;
    else    Extract paragraph initial and last sentences;        // as suggested in DRPA
}
Scoring and Discrimination;
```

Conclusions on the Classroom Experiment

The title weighting and the hot spot method provide a guide for DRPA to distill representative phrases more flexibly and accurately. More than that, in the assaying process, the technique can keep recursive “reading” and “scanning” the document for several times when judging whether the distillation needs to enroll into the second round or third round. We can see that the refined DRPA has the features of prioritization and the hierarchical structure. It is “intelligent” enough to ***make decision conditionally***. (Please refer to the refined steps in the previous subsection).

Although there is no further experiment to test on other patterns for discrimination, it is believed that the performance is satisfactory even in other domains or patterns because the suggested approach for improving DRPA is to consider and dependent more on natural language understanding.

CONCLUSIONS

Train students the discernment ability is an important task in distance education. To improve the training effects, we have proposed a new method, called DRPA, to discriminate courseware document based on natural language understanding and Representative Phrase Assaying. Our main works includes:

- Suggests the web courseware on controversial social issues with different viewpoints.
- To construct the courseware for Discernment Ability, a method called DRPA (Discriminating via Representative Phrase Assaying) is presented with five-algorithms, i.e. algorithms for extracting representative phrases, algorithms for calculating characteristic array, algorithms for determining the threshold array, algorithms for objective judgment, and algorithms subjective judgment.
- A new concept called Gymnastics Threshold is proposed and proved to be more accuracy than transitional threshold.
- Extensive experiment algorithms are given. Our experimental results show that the algorithms are efficient. As a web courseware can be rather complex, much further work needs to be done. We are currently investigating issues on improving the accuracy of the evaluation and mining the threshold array by clustering technique.

Training students on the discernment ability is a new issue in Web education. This is just a beginning work. A lot of work remains to be done, including how to manage the examination for discernment ability training, and how to collect the Web documents with different viewpoints, etc..

ACKNOWLEDGEMENT

This project is in part supported by a CERG grant from the Research Grants Council of Hong Kong (RGC Reference Number: CPHK 77/95E), the National Science Foundation of China Grant #60073046,and the Specialized Research Fund for the Doctoral Program of Higher Education(SRFDP）#20020610007.

REFERENCES

Tang C., Lau R.W.H., Li Q., Yin H., Li T., and Kilis D. (2000). Personalized Courseware Construction Based on Web Data Mining, WISE'00 Worshop on Web-based Education and Learning, 204-211.

Tang C., Lau R.W.H., Yin H., Li Q., Lu Y., Yu Z., Xiang L., and Zhang T. (1999). Discovering Tendency Association Between Objects with Relaxed Periodicity and its Application in Seismology, Proceedings of ICSC '99, LNCS Vol. 1749, 51-62.

*Tang C., Yu Z., You Z., Zhang T., and Yang L. (2000). Mine the Quasi-Periodicity From Web Data, The Journal of Computer, **23**(1):52-59.*

*Riloff E. and Lehnert W. (1994). Information Extraction as a Basis for High-precision Text Classification, ACM Transactions on Information Systems, **12**(3):296-333.*

Mannila H. and Toivonen H. (1999). Discovering Generalized Episodes Using Minimal Occurrence, Proceedings of the International Conference on Knowledge Discovery and Data Mining.

Jiang M., Tseng S., and Tsai C. (1999). Discovering Structure from Document Databases, Proceedings of PAKDD'99, 169-173.

Fayyad U. and Piatetsky G. (1996), Advanced in Knowledge Discover and Data Mining (Eds), AAAI Press and MIT Press, 1-5.

Zuo J, Tang C., and Zhang T. (2002). Mining Predicate Association Rule by Gene Expression Programming, WAIM02, International Conference for Web Information Age, LNCS Vol. 2419, 92-103.

*Yu Z., Tang C. and Zhang T. (2000), The Grammar Analysis Strategy for Machine Translation Systems, Journal of Micro and Mini Computer Systems, **21**(3) 316-318.*

*Damerau C. and Weiss F. (1994), Automated learning of decision rules for text categorization, ACM Transactions on Information Systems, **12**(3):233~251.*

*Li T, Tang C., Zuo J., and Zhang T. (2001). Web Document Filtering Technique Based on Natural Language Understanding, International Journal Computer Processing of Oriental Language, **14**(3):279-291.*

Tang C, Li Q., Lau R.W.H., and Huang X. (2003). Supporting Practices in Web-based Learning, ICWL03 (First International Conference on Advances in Web-Based Learning), LNCS Vol. 2436, 300-312.

The authors' biography

TANG Changjie *received his MSc. from Department of Mathematics, Sichuan University in 1981. His current interests are in Web data mining. He has published 8 books and more than one hundred research papers in journals and international conferences, such as FODO, IFIP, SIGMOD, DASFAA, ICSC, TCS (Theoretical Computer Science) ,LNCS, JOS, JCST, SC (Science of China). He was a PC member of VLDB'97, DASFAA'99, ICSC'99, WAIM2000-04,* **and DASFAA*2001-2004*.** *He is a Professor of the Computer Department of the Sichuna University, and a vice director of the Database Society of the Chinese Computer Federation. (Email:chjtang@vip.sina.com, or tangchangjie@cs.scu.edu.cn)*

Rynson W.H. Lau *received a B.Sc. first-class honors degree in Computer Systems Engineering in 1988 from the University of Kent at Canterbury, England, and a Ph.D. degree in Computer Science in 1992 from the University of Cambridge, England. He is currently an associate professor at the City University of Hong Kong. Prior to joining the university in 1998, he taught at the Hong Kong Polytechnic University. From 1992 to 1993, he worked at the University of York, England, on a defense project on image processing. Rynson Lau's research interests include computer graphics, virtual reality and multimedia systems. (Email: Rynson.Lau@cityu.edu.hk)*

Qing Li *received his BEng. degree from Hunan University (Changsha, China), MSc and PhD degrees from the University of Souther California (Los Angeles, USA), all in computer science. He is currently an Associate Professor at the City University of Hong Kong, as well as a Guest Professor of the Zhejiang University, and an Adjunct Professor of the Hunan University. His research interests include database modeling, multimedia retrieval and management, and e-learning systems. Dr Li has published over 150 papers in technical journals and international conferences in these areas, and is actively involved in the research community by serving as a guest and asssocaite editor to several technical journals, programme committee chair/co-chair, and as an organizer/co-organizer of major international conferences. Currently he serves as a councillor of the Database Society of Chinese Computer Federation, and as a Steering Committee member of the international WISE Society. (Email: itqli@cityu.edu.hk)*

Jean W.H. Poon *received the Bsc from City University of Hong Kong in 2002 and M.A. degree from Chinese University of Hong Kong in 2004. She is currently a teacher of a secondary school in Hong Kong. (E-mail: waihan_poon@yahoo.com.hk)*

Zhang Tianqing *received the B.E. and M.E degree in computer science from Sichuan University, China, in 1993 and 1996 respectively. He is currently a Ph D candidate. in Computer Science, Sichuan University. His main research interests are database system. (Email:zhangtianqing@cs.scu.edu.cn)*