A Design Framework for Online Discussion Speech Act Analysis and Situation Profiling

Jia Li[1] Xuan Liu[1] Zhigao Chen[1] Pengzhu Zhang[2]

[1]School of Business, East China University of Science and Technology,
Shanghai, P.R. China
Email: jiali@ecust.edu.cn, xuanliu@ecust.edu.cn, zgchen@ecust.edu.cn
[2]Antai School of Economics and Management, Shanghai Jiaotong University,
Shanghai, P.R. China
Email: pzzhang@sjtu.edu.cn

Abstract

In this research, we developed a design framework for systems supporting online discussion speech act analysis and situation profiling. The framework advocates the development of systems that support all three facets of online discussion situation profiling: snapshot, duration, and people. The framework also provides guidelines for the creation of speech act category, the choice of appropriate classification algorithms, features, and feature selection techniques necessary to effectively identify the roles each message plays, and the topic categorization of discussion text.

Keywords: situation profiling, speech act analysis, online discussion, design framework

1. Introduction

Discussion situation profiling is considered as a key step toward opening the "black box" [1][2] of computer mediated group decision making. For example, by identifying the percentage of participants who contributed to the alternative, facilitators can assess if majority members are involved and decide whether further calling for participation is necessary. Another example is identifying the consensus state of the alternatives, based on which users can understand to what extent the participants has reached agreement on each alternative and decide if the group should put more effort on the unsolved ones.

However, systematical profiling group discussion situation can be quite difficult under the condition of information overload. As computer-supported groups are confronted with larger numbers of ideas and supporting comments to organize and evaluate, they may experience information overload [3]. Group Support Systems allow simultaneous, immediate entry and storage of comments, and enables people to enter comments as they think of them, without having to wait their turn [4-6]. As a result, often as many as several hundred comments can he generated by a group of l0-20 meeting participants during a typical one-hour electronic meeting session. Thus the discussion situation identification task, e.g., identifying the relationships among messages, identifying the solved/unsolved issue, identifying mature/immature discussions, identifying the temporal change of discussion state, identifying the statistic profile of participants etc., becomes big challenges for meeting participants under the condition of information overload. Users have got to remember all aspects of discussion information by themselves before they can make a correct assessment. When discussion state identification becomes a bottleneck, as is often the case, it counteracts the productivity gains and reduces the satisfaction of meeting participants.

As human users may have difficulty in profiling discussion situation, a system that automatically analyzes discussion text and profiles discussion situation is highly desirable. This paper thus describes a design framework inclusion of online discussion speech act analysis and situation profiling. Following Walls et al.'s model [7], we present the kernel theory, meta-requirements, meta-design and testable hypothesis in the rest of this paper.

2. A Design Framework for Online Discussion Situation Profiling

According to the design science paradigm, design is a product and a process [7, 8]. The design product is the set of requirements and necessary design characteristics that should guide IT artifact construction. Walls et al. [7] presented a model for the formulation of information systems design theories (ISDTs). Their model incorporates four components guiding the design product aspect of an ISDT. These include the kernel theories, meta-requirements, meta-design, and testable hypotheses. The kernel theories govern meta-requirements for the design product. The meta-design is anticipated to fulfill these meta-requirements by providing detailed specifications for the class of IT artifacts addressed by the design product. Testable hypotheses are used to evaluate how well the meta-design satisfies meta-requirements.

Using Walls et al.'s model, we propose a design framework for online discussion speech act analysis and situation profiling systems. Employing speech act theory, argumentation theory, and situation awareness theory as our kernel theory, we propose meta-requirements and a meta-design necessary to support speech act analysis and situation profiling. We also present hypotheses intended to evaluate how well the meta-design satisfies our meta-requirements. The following sections elaborate on the components of our design framework.

3. Kernel Theories

3.1. Speech Act Theory

Speech act is a technical term in linguistics and the philosophy of language. Speech acts can be analyzed on three levels: A locutionary act, the performance of an utterance: the actual utterance and its ostensible meaning, comprising phonetic, phatic and rhetic acts corresponding to the verbal, syntactic and semantic aspects of any meaningful utterance; an illocutionary act: the semantic 'illocutionary force' of the utterance, thus its real, intended meaning; and in certain cases a further perlocutionary act: its actual effect, such as persuading, convincing, scaring, enlightening, inspiring, or otherwise getting someone to do or realize something, whether intended or not [9].

The concept of an illocutionary act is central to the concept of a speech act. According to Austin's preliminary informal description, the idea of an "illocutionary act" can be captured by emphasizing that "by saying something, we do something", as when someone orders someone else to go by saying "Go!", or when a minister joins two people in marriage saying, "I now pronounce you husband and wife."

Speech act theory been applied to many domains such as healthcare. Speech Act analysis allows for a useful understanding of the status of a negotiation between (for instance) a health care provider and a patient independent of any well-accepted credible and comprehensive understanding of a disease process as it might apply to that patient. For this reason, systems which track the status of promises and rejected-proposals and accepted-promises can help us to understand the situations in which (human or computer) agents find themselves as they attempt to fulfill roles involving other agents, and such systems can facilitate both human and human-computer systems in achieving role-associated goals.

In this research, computer mediated discussion are considered a special case of human conversation, and each utterance is classified according to speech act categories such as question, answer, issue and acknowledgement. We believe that speech act analysis could be helpful to understand discussion situation in complicated online discussion.

3.2. Argumentation Theory

Argumentation is a verbal and social activity of reason aimed at increasing (or decreasing) the acceptability of a controversial standpoint for the listener or reader, by putting forward a constellation of propositions intended to justify (or refute) the standpoint before a rational judge [10]. Argumentation is a verbal activity, most often in an ordinary language. In argumentation people use words and sentences to argue, to state or to deny etc. Furthermore, argumentation is a social activity, which in principle is directed to other people. Argumentation is also an activity of reason, when people put forward their arguments in argumentation they place their considerations within the realm of reason. Argumentation is always related to a standpoint. An opinion itself is not enough; arguments are needed when people differ on a standpoint. Finally, the goal of argumentation is to justify one’s standpoint or to refute someone else’s.

Toulmin uses a model of argumentation for his “uses for argument” [11]. Brockriede and Ehninger [12] refer to Toulmin and describe an argument as “movement from accepted data, through a warrant, to a claim.” Toulmin’s argument model is presented in Figure 1. The three main components in argumentation theory include data, warrant, and claim. Data refers to the facts or opinions of evidence. Claim refers to the conclusion. Warrant is the “leap” which advances data to a claim [12]. Specifically Toulmin [11] says that warrant is “incidental and explanatory, its task being simply to register explicitly the legitimacy of the step involved and to refer it back to the larger class of steps whose legitimacy is being proposed.”

The second set of components is not necessary, but may be present. These additional three components include backing, rebuttal, and qualifier. Backing refers to the evidence or support for assumptions in the warrant. Rebuttal recognizes the conditions under which the claim will not be true. Finally, qualifier is the probability or level of confidence of the claim.

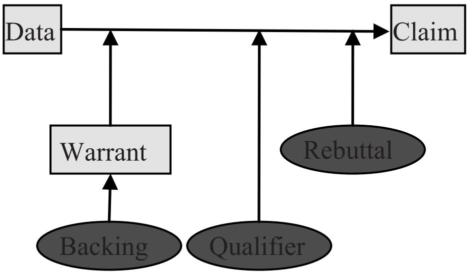

Fig. 1: Toulmin’s argument model [11].

3.3. Situation awareness theory

Situation awareness, or SA, is the perception of environmental elements within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future. SA involves being aware of what is happening around you to understand how information, events, and your own actions will impact your goals and objectives, both now and in the near future. Lacking SA or having inadequate SA has been identified as one of the primary factors in accidents attributed to human error [13-15]. Thus, SA is especially important in work domains where the information flow can be quite high and poor decisions may lead to serious consequences (e.g., piloting an airplane, functioning as a soldier, or treating critically ill or injured patients).

The most common theoretical framework of SA is provided by Dr. Mica Endsley [16]. Endsley's model illustrates three stages or steps of SA formation: perception, comprehension, and projection.

Perception (Level 1 SA): The first step in achieving SA is to perceive the status, attributes, and dynamics of relevant elements in the environment. Thus, Level 1 SA, the most basic level of SA, involves the processes of monitoring, cue detection, and simple recognition, which lead to an awareness of multiple situational elements (objects, events, people, systems, environmental factors) and their current states (locations, conditions, modes, actions).

Comprehension (Level 2 SA): The next step in SA formation involves a synthesis of disjointed Level 1 SA elements through the processes of pattern recognition, interpretation, and evaluation. Level 2 SA requires integrating this information to understand how it will impact upon the individual's goals and objectives. This includes developing a comprehensive picture of the world, or of that portion of the world of concern to the individual.

Projection (Level 3 SA): The third and highest level of SA involves the ability to project the future actions of the elements in the environment. Level 3 SA is achieved through knowledge of the status and dynamics of the elements and comprehension of the situation (Levels 1 and 2 SA), and then extrapolating this information forward in time to determine how it will affect future states of the operational environment.

SA also involves a temporal component. Time is an important concept in SA, as SA is a dynamic construct, changing at a tempo dictated by the actions of individuals, task characteristics, and the surrounding environment. As new inputs enter the system, the individual incorporates them into this mental representation, making changes as necessary in plans and actions in order to achieve the desired goals.

4. Meta-Requirements

Effective profiling of group discussion situation entails consideration of representing three facets: snapshot profiling, duration profiling and people profile [17].

4.1. Situation profiling

(1) Active topics vs. inactive topics

The active topics are live topics which are under discussion and ready to receive coming comment from users. The inactive topics are fading topics or closed topics which are out of discussion focus. Inactive topics receive few or non comments from users and are not expected to receive a few in the future.

Discussion group usually explores the debate space by move topics from one subcategory to another. Tracing activeness of topic provides a mean to understand how the debate space is explored.

(2) Majority topics vs. minority topics

Different topics may receive different levels of participation. Representativeness is a metric measuring to what degree the discussion result can represent the opinion of majority. The majority topics attract most of group members to participate, so the discussion result can be considered as valid in terms of majority rule. On the other side, the minority topics attract only few group members to participate. The discussion result of minority topics should be carefully adopted even if the discussion is flourishing.

(3) Hot topics vs. rare topics

Hot topics are ones receiving many comments, while rare topics are ones receiving few comments. Hot topics attract more comments than rare topics because: 1) the content hot topics are more interesting; 2) the hot topics with many comments are more likely to receive more comments than rare topics, i.e., the rule of “the rich get richer”.

Although hot topics are always the focus of discussion, rare topics are also valuable for group discussion. Rare topics usually corresponds to immature discussion, accompany with low discussion depth and width. The rare topics are important because they may contain insightful opinion but overlooked.

(4) Agreed topic vs. conflicting topic

Consensus measures to what degree the discussant has agreed on. Since everyone’s opinion is encouraged and valued, group consensus is a critical factor of group decision making. Agreed topics are ones on which most users have positively or negatively agreed, while controversial topics are ones on which the group opinion are under conflicting.

Consensus is a key concept in group decision making, while in most cases group members are seeking agreement. One of the tendency typically found in group interaction is “not changing the subject” [18]. By identifying topics to which the discussion are mature and agreed, the facilitator can ask group member jump out of the fixed frame of those topics and try to explore more space of solution to improve group effectiveness and efficiency.

It is notable that it doesn’t make any sense to measure consensus if the topic is discussed by only few group members (low representativeness).

(5) Critical atmosphere vs. supportive atmosphere

Ideally, the discussion atmosphere should be a balance of critical and supportive. Mason [19] suggested that dialectical inquiry that’s both critical and constructive should lead to higher quality solutions. Identifying meeting atmosphere is important for facilitator and group members. By providing them with situation of meeting atmosphere, group members may discuss how well their group is functioning and how group processes may be improved [20]. During these discussions group members may be stimulated to adopt more critical or exploratory group-norms.

4.2. Group profiling

(1) Supporters vs. opponents

Profiling supports and opponents is important for further analysis of conflicting topics. Conflicting topics are usually accompanied with two sides: the supporters and opponents. Identifying the supporters and opponents are helpful for an intuitive understanding the balance of two sides.

Profiling supporters and opponents is also helpful for identifying hidden group, a sub group hidden in a bigger group with common interest and attitude. A critical cue to identify a hidden group is that they share common attitude in different sub-categories of discussion.

(2) Contributors vs. lurkers

Contributors are those who post many messages, while lurkers are those who post few or none messages. Previous studies have found that the number of messages sent out by participants indicates their attitude toward the community [33]. An individual active in discussion may not be the most knowledgeable person, but he or she is probably willing to contribute to the group.

(3) Attractors vs. ordinaries

Different messages may have different effect on attracting reply post. Some messages have higher reply-num than others. The reply-num may indicate the quality of post, i.e., higher reply-num indicating higher quality of post.

The situation is the same for person involved in the discussion. Messages posted by some person attract more post, while by others less. Attractors are such persons whose messages attract more posts than ordinaries. Identifying attractors is important because attractive is considered as a character of leader in the group.

(4) Authorities vs. ordinaries

Different messages may have different effect on getting positive reply. Some messages have better feedback than others. Feedback from other users may be another indicator to the quality of message, i.e., positive feedback indicating higher quality of post.

The situation is the same for person involved in the discussion. Messages posted by some persons attract more positive reply, while by others less. Authorities are such persons whose messages attract more positive replies than ordinaries. Identifying authorities is also important because positive feedback is considered as another character of leader in the group.

(5) Balanced participators vs. biased participators

Ideally, user should cover all parts of the discussion proportionally. However, user usually focuses on some while ignores the others. Previous studies show that group discussion can suffer from cognitive inertia, the tendency of group discussion to focus on a few lines of thought in one subcategory [18]. As group members interact, they may consciously or unconsciously adopt behavior norms. These norms or structures can constrain behavior [21]. One of the structures typically found in group interaction is “not changing the subject” [18]. By identifying users that focus only on limited sub-topics of the discussion and encouraging their balanced participation, one can expect more space of debate explored which leads to improved performance.

4.3. Individual profiling

(1) Expertise

Group members with different knowledge and experience usually have different type of expertise. Expertise consists of characteristics, skills and knowledge of a person (that is, expert), which distinguish experts from novices and less experienced people. Finding the experts in an organization and help people knows how to reach the experts is an important task in knowledge management. The subtopics where users often receive positive replies can be considered as sub-categories of interest.

(2) Interest

Interest is a state of curiosity or concern about or attention to something. Interest and expertise might not be identical in many cases. The expertise corresponds to a status of knowledgeable, while interest corresponds to curiosity or concern. The subtopics that users are actively involved can be considered as sub-categories of interest.

(3) Criticalness

Criticalness may reveal the user's characteristic and preference to the discussion. By judging user's criticalness, we can find different types of users: some may be good at raising recommendations and assumptions, while others good at critiques of single sets of recommendations and assumptions. Over critical users are usually unpopular to the group norms but their challenges are valuable to critical thinking.

5. Meta-Design

While meta-requirements are derived from the kernel theories, the objective of the meta-design is to introduce a class of artifacts hypothesized to meet the meta-requirements [7]. Critical elements of speech act analysis system are speech act category, classification algorithm, features, and feature selection techniques. For discussion situation profiling system, text categorization and group profiling model are critical designs.

5.1. Speech act taxonomy design

Speech act taxonomy is different in different domains. Speech Act taxonomy must compromise between two factors. First, the definitions of SA tags must be clear enough in order to be easily separable. If they are not, agreement between human taggers will be low. On the other hand it is efficient to define a reusable taxonomy, which is general enough to be applicable to many different problems.

There seems to be little agreement on how exactly to achieve the compromise. The most popular taxonomy, initially designed to be universal, is DAMSL. Other taxonomies have been developed for some corpora, like CallHome or VerbMobil, and have later on gained popularity. A possible way to create the speech act category is to start from one or two existing taxonomies and apply some modifications to adapt to the new domain.

A speech act category can be considered as reliable only if the inter-rater reliability is high enough. The options for determining inter-rater reliability include joint-probability of agreement, Cohen's kappa and the related Fleiss' kappa, inter-rater correlation, concordance correlation coefficient and intra-class correlation.

5.2. Speech act classification algorithm

For the contemporary techniques, both the utterance and the speech act have to be encoded in order to be used as the classification algorithm's input and output respectively. A common way to encode an utterance is to describe its words with features. The output (the speech act) is encoded as a nominal feature.

Transformation-based learning (TBL) was introduced by Brill (1993). It is based on a set of rules, which are applied consecutively to the data, changing some tags into other ones. The rules are controlled by preset templates; the most common ones are of the type "if current tag is A, it is preceded by tag B and/or the word C is present in one of the preceding N utterances, change the current tag to D".

Rules are composed in a supervised manner. Having a marked training corpus, all possible rules are generated from the templates, after which the rules are selected iteratively: the rule bringing the biggest improvement to the precision is selected on each iteration. The process is continued until one of the stopping criteria is met; the most common is that no improvement is brought by applying any rule.

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. A support vector machine constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training datapoints of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

5.3. Features for speech act classification

Text features are linguistic attributes used to represent various information types. They can be classified into two broad categories: language resources and processing resources [22]. Language resources are data-only resources such as lexicons, thesauruses, and word lists. These self-standing features exist independent of their application context and provide powerful discriminatory potential. However, language resource construction is often manual, and features may be less generalizable across information types [23].

Processing resources require algorithms for computation. Parts-of-speech tags, n-grams, statistical features (e.g., average word length), and bag-of-words are all examples of processing resources. The majority of processing resource features are context-dependent; they change according to the text corpus. However, the extraction procedures remain constant, making processing resources highly generalizable across information types. Consequently, features such as bag-of-words, part-of-speech tags, and n-grams are helpful to recognize the speech act of each message.

5.4. Feature Selection Techniques for speech act classification

Two categories of feature selection techniques commonly applied to text are ranking and projection based methods [24]. Ranking techniques rank attributes based on some heuristic [25]. Examples include information gain, chi-squared, and Pearson’s correlation coefficient [26, 27]. Projection methods are transformation based techniques that utilize dimensionality reduction [28]. Examples are principal component analysis (PCA), multidimensional scaling (MDS), and self-organizing map. Ranking and projection based methods each have their advantages and disadvantages. Ranking methods offer greater explanatory potential than projection methods since they preserve the original feature set and simply rank/sort attributes. Ranking methods also offer simplicity and scalability. However, they typically consider only an individual feature’s predictive power; resulting in the potential loss of information stemming from feature interactions. Projection methods are highly robust against noise, making them useful for text analysis. However, the transformation process from original features to projections can also diminish explanatory potential.

5.5. Text Categorization

The text categorization consists of indexing and categorization. Indexer automatically represents a document with a vector of terms [29]. Many tools like the Arizona Noun Phraser (AZNP) were created to identifying nouns from sentence. With help of noun parser, indexer represents each message with key phrases identified. Categorizer automatically categorizes the discussion content and identifies subtopics. Kohonen's self-organizing maps (SOM) appears to be a promising algorithm for organizing large volumes of information. SOM was first proposed by Kohonen, who based his neural network on the associative neural properties of the brain [30]. The network consists of an input layer and an output layer. The number of the input nodes equals the number of attributes associated with the input. After all of the input is processed, the result is a spatial representation of the input data, organized into clusters of similar regions.

The output of text categorization is helpful to identify the user's expertise and interest. Expertise usually corresponds to positive feedback from others. A sub-category is considered as an expertise topic for a user if he or she always receives positive reply from other group members. Interest corresponds to high participation. A sub-category is considered as an interest topic for a user if he or she is active in contributing to that topic.

5.6. Group profiling model

The group profiling model provides operational definitions of meta-requirements mentioned in section 4 using the results derived from speech act analysis. Following that model, a systematical profiling of discussion situation satisfying meta-requirements can be build. For more details about the group profiling model, see [17].

6. Testable Hypotheses

Testable hypotheses are intended to assess whether the meta-design satisfies meta-requirements [7]. For the proposed design framework, this entails evaluating the meta-design's ability to accurately represent information types associated with the three meta-functions, as outlined in the meta-requirements. In this research, we test the effectiveness of situation profiling by comparing to the result from human experts.

7. Conclusion and Future Work

In this research, we developed a design framework for systems supporting online discussion speech act analysis and situation profiling using Walls et al.'s [7] model. The framework advocates the development of systems that support all three facets of online discussion situation profiling: snapshot, duration, and people. The framework also provides guidelines for the creation of speech act category, the choice of appropriate classification algorithms, features, and feature selection techniques necessary to effectively identify the roles each message plays, and the topic categorization of discussion text.

This research is part of our ongoing project dedicated to profile discussion situation and provide necessary interventions to benefit the group automatically or semi-automatically. Our future work includes designing and evaluating a system in light of the design framework proposed in this research.

Acknowledgement

This research was supported by the National Natural Science Foundation of China under Grant 71001038.

References

[1] Dennis A R, Garfield M J. The adoption and use of GSS in project teams:

Toward more participative processes and outcomes [J]. *MIS Quarterly*, 2003, 27(2): 289-323.

[2] Limayem M, Banerjee P, Ma L. Impact of GDSS: opening the black box [J]. *Decision Support Systems*, 2006, 42(2): 945-957.

[3] Grise M, Gallupe R B. Information overload: addressing the productivity paradox in Face-to-Face electronic meetings [J]. *Journal of Management Information Systems,* 2000, 16(3): 157-185.

[4] Gallupe R B, Bastianutti L M, Cooper W H. Unblocking brainstorms [J]. *Journal of Applied Psychology*, 1991, 76(1): 137-142.

[5] Gallupe R B, Cooper W H, Grisé M L, Bastianutti L M. Blocking electronic brainstorms [J]. *Journal of Applied Psychology*, 1994, 79(1): 77-86.

[6] Gallupe R B, Dennis A R, Cooper W H, Valacich J S, Bastianutti L M, Nunamaker Jr J F. Electronic brainstorming and group size [J]. *Academy of Management Journal*, 1992, 35(2): 350-369.

[7] Walls J G, Widmeyer G R, El Sawy O A. Building an information system design theory for vigilant EIS [J]. *Information Systems Research*, 1992, 3(1): 36-59.

[8] Hevner A R, March S T, Park J, Ram S. Design science in information systems research [J]. *MIS Quarterly*, 2004, 28(1): 75-105.

[9] Austin J L. How to do things with words [M]. Cambridge: Harvard University Press, 1962.

[10] Van Eemeren F H, Grootendorst R, Henkemans F S. Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments [M]. Lawrence Erlbaum Ass, 1996.

[11] Toulmin S. The Uses of Argument [M]. Cambridge, MA: Cambridge University Press, 1958.

[12] Brockriede W, Ehninger D. Toulmin on argument: An interpretation and application [J]. *Quarterly Journal of Speech*, 1960, 46(1): 44-53.

[13] Hartel C E, Smith K, Prince C. Defining aircrew coordination: Searching mishaps for meaning [C]. *The 6th International Symposium on Aviation Psychology*. Columbus, OH, 1991.

[14] Merket D C, Bergondy M L, Salas E. Making sense out of team performance errors in military aviation environments [J]. T*ransportation Human Factors*, 1999, 1(3): 231-242.

[15] Nullmeyer R T, Stella L C D, Montijo G A, Harden S W. Human factors in Air Force flight mishaps: Implications for change [C]. Proceedings of t*he 27th Annual Interservice/Industry Training, Simulation, and Education Conference.* Arlington, VA, National Training Systems Association, 2005.

[16] Endsley M R. Toward a theory of situation awareness in dynamic systems [J]. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 1995, 37(1): 32-64.

[17] Li J, Liu X, Zhang P Z. Measuring situation awareness in computer mediated discussion [C]. *Third International Symposium on Intelligent Ubiquitous Computing and Education. Beijing*, IEEE Press, 2010.

[18] Lamm H, Trommsdorff G. Group versus individual performance on tasks requiring ideational proficiency (brainstorming): a review [J]. *European Journal of Social Psychology*, 1973, 3(4): 361-387.

[19] Mason R O. A dialectical approach to strategic planning [J]. *Management Science*, 1969, 15(8): 403-414.

[20] Yager S, Johnson R T, Johnson D W, Snider B. The impact of group

processing on achievement in cooperative learning groups [J]. *Journal of Social Psychology,* 1986, 126(3): 389-397.

[21] Giddens A. The constitution of society: outline of the theory of structure [M]. Berkeley, CA: University of California Press, 1984.

[22] Cunningham H. GATE, a general architecture for text engineering [J]. *Computers and the Humanities,* 2002, 36(2): 223-254.

[23] Pang B, Lee L, Vaithyanathan S. Thumbs up? sentiment classification using machine learning techniques [C]. Proceedings of *the ACL Conference on Empirical Methods in Natural Language Processing*. Philadelphia, PA, Association for Computational Linguistics, 2002.

[24] Guyon I, Elisseeff A. An introduction to variable and feature selection [J]. *The Journal of Machine Learning Research*, 2003, 3: 1157-1182.

[25] Hearst M A. Untangling text data mining [C]. Proceedings of *the 37th Annual Meeting of the Association for Computational Linguistics. College Park,* MD, Association for Computational Linguistics, 1999.

[26] Forman G. An extensive empirical study of feature selection metrics for text classification [J]. *The Journal of Machine Learning Research*, 2003, 3: 1289-1305.

[27] Koppel M, Schler J. Exploiting stylistic idiosyncrasies for authorship attribution [C]. Proceedings of *IJCAI Workshop on Computational Approaches to Style Analysis and Synthesis*. Acapulco, Mexico, 2003.

[28] Huang S, Ward M O, Rundensteiner E A. Exploration of dimensionality reduction for text visualization [C]. Proceedings of *The Third International Conference on Coordinated and Multiple Views in Exploratory Visualization*. London, 2005.

[29] Chowdhury G G. Introduction to modern information retrieval [M]. McGraw Hill Computer Science Series, 2004.

[30] Kohonen T. Self-Organization Maps [J]. *Proc of IEEE*, 1990, 78(9): 1464-1480.