Proc. of The Ninth Int. Symp. on Artificial Life and Robotics (AROB 9th '04)
Beppu, Oita, Japan, 28-30 January, 2004

Toward the Consistent Simulation of Narrative Discourse Based on Narratology

Takashi Ogata
Department of Medicine and Engineering, Yamanashi University
Yamanashi, 400-8511 Japan
ogata@csi.yamanashi.ac.jp

Abstract

We discuss aspects of the narrative rhetoric based on narratology and a direction of their integration into a consistent narrative generation system. In traditional narratology, the typology of narrative rhetoric, the application to texts interpretation, implications to ideological and social thinking of texts interpretation, and so on are mainly treated. But we propose a new direction of narratology to be oriented toward so called an operational approach based on computational modeling and simulation of narrative generation process. In this paper, at first, we explain the computational re-definitions of four kinds of narrative rhetoric about narrative discourse aspect, and show a discourse transformation process using methods partially implemented.

1 Introduction: From Typological/ Interpretive Narratology to Operational Narratology

By the integration of cognitive and computational techniques and theories about narrative generation & understanding and literary knowledge based on literary theories and narratology, we have researched about, so to speak, a narrative generation mechanism supported by literary knowledge [1]. A central interest in traditional literary studies was the analysis and interpretation of literary texts and relating literary phenomena, but our approach aims at the conversion from such analytic and interpretive direction to literary theories for production and creation.

By adopting and expanding the basic idea of narratology, we divide a narrative generation process into three aspects of story, discourse, and representation or narration. Story means a temporal sequence of events to be narrated and a kind of narrative potential structure. Propp [2] proposed an influential literary theory about narrative potential structure and transformation based on it. As a starting point, we applied his theory to develop experimental story generation systems [3, 4]. On the other hand, narrative discourse is an aspect of how to narrate a story in various ways, and directly reflects the structure of surface text.

In this paper, we attempt a consideration toward the mechanism of a consistent process of the narrative discourse aspect by using Genette's theory [5] as a starting point. About this aspect in narratology (structural study on the aspect of narrative discourse is thought as narratology in the narrow sense by researchers), we have studied along following research process; (1) the computational re-definition of Genette's discourse elements. Namely, we have re-defined Genette's typological description as more precise and operational rhetorical techniques introducing the methods of computer modeling and simulation. (2) The integration of each discourse aspect and the composition of a consistent discourse simulation system. This paper considers about the approach to the second problem.

2 Narrative Techniqeus for Discourse Processing

Genette's discourse theory is composed of following three aspects; time or tense, mood, and voice. Time or tense means the set of temporal relations (speed or duration, order, and frequency) between story (narrative contents) and discourse. Mood, next major category, is the set of modalities, distance and perspective (point of view or focalization) regulating narrative information. Voice, last major category, treats the temporal and spatial relations between a narrator's narrative action and a narrated text. We have done the precise analysis and the implementation of experimental simulation systems of order & duration in time and distance & perspective in mood. Their detailed concepts found as operational narrative techniques are explained in next sections.

2.1 Temporal order's techniques

Order shows the order in which events occur and the order in which they are recounted. Genette has made a precise classification of the temporal order, and we have used it in our system by re-defining from the viewpoint of computation. We have called the temporal order processing based on Genette *structural discourse techniques of time*.

For example, structural transformation of temporal order is formally made by the substitution of some events in a sequence of events. But, a reader can not sometimes understand the existence of order transformation. Therefore, in real narrative texts, techniques that, so to speak, compensate for the structural discourse techniques are frequently used. For example, by the action of "recall" in a character, a temporal changing of order occurs, and simultaneously the recognition of a reader on introducing of a structural discourse techniques of time is helped. *The techniques of operational discourse of time* are the techniques in the level which makes a reader recognize the use of temporal structure's changing. We have proposed a computational modeling of time in

which blends these two types of narrative techniques [6].

Operational discourse techniques of time are operated by two kinds of subjects; character and narrator. It also has two directions; introducing or inserting the past into the present and introducing or inserting the future into the present. It has various sub-categories, namely concrete operations. Originally, we acquired these operations through the analysis of movie films, but here classify the operations into more abstract groups of operations; (1) Introducing the past and the future into the present by character-action: For example, the action in which a character recalls or dreams a past event is inserted into the present. As other ways, transmission is also used. There are predictions, premonition, planning, and so on in ways for introducing the future into the present. (2) Introducing the past and the future into the present by character-voice: In the voice, there are monologue and dialogue. (3) Introducing the past and the future into the present by narrator-action: In all narratives, the concept of narrator that comprehends characters exists. A narrator directly can insert a scene of the past or the future which a character does not know. (4) Introducing the past and the future into the present by narrator-voice: Similarly, a narrator introduce the past or the future into the present by narration. (5) Introducing the past and the future into the present by objects: This is the case that the past or future events are indicated by the existence of objects such as newspaper, ring, and scar.

2.2 Perspective's techniques

Perspective is the perceptual position in terms of which the narrated situations and events are rendered. It is divided into three types; zero perspective, internal perspective, and external perspective.

We define perspective as the problem of selection and removing in a story's information, and make a simple mechanism that transforms or edits the conceptual expression of a given story to different conceptual expressions based on different types of perspectives [7]. We think simply that the processing of perspective is the problem of deciding the range of information inside a story to be transformed to narrative discourse. We divide Genette's classification of perspective into two categories; first is a group of basic perspectives and second is a group of applicable ones. The applicable perspectives can be defined by the combinations of basic perspectives so is strongly related with strategic knowledge.

Basic perspectives have three sub types; zero perspective, internal fixed perspective, and external perspective. Zero perspective is he type which can narrate all information in a story. Internal fixed perspective is the type which narrates narrative information about a certain character and information the character perceives such as mental state and the stream of consciousness. External perspective is the type which narrates only external aspects about characters.

Applicable perspectives have internal variable perspectives and internal multiple perspectives which are two applications of the internal fixed perspective. Moreover, as another type's category, alteration means a temporary transgression to the perspective adopted consistently, and has paralipsis and paralepsis. The internal variable perspective is the type which narrates a story while changing the character chosen as the subject of a perspective (focalized character). Internal multiple perspective is the type which narrates a same event using different characters' internal fixed perspectives. The paralipsis is the technique that does not intentionally (perhaps) narrate the narrative information which should be narrated, and the paralepsis is one that dare narrate the narrative information which should be overlooked.

We implemented a first experimental system by Allegro Common Lisp which transformed variously conceptual data expressing a story based on the specification of perspectives. A user gives a sequence of events, the information related to their events, and a particular type of perspective. And then, the system outputs a sequence of transformed events. The conceptual expression of each event in a story has an event concept as first item and pairs of some slots like actor, object, place, and their values. Each element in an event concept is linked to a particular element in persona-frames, object-frames, and place-frames. Each slot in a persona-frame corresponds to name, sex, external-aspect, internal-aspect and perception. The perception slot stores an event that a character currently perceives. Object-frame and place-frame respectively have a name slot and an external-feature slot.

2.3 Distance's techniques

The concept of distance is originated with Plato. Plato called the state that narrates in the name of a poet (author) itself a story "diegesis" (pure narrative discourse), and the state that a story is narrated without the intervention of any poets or authors "mimesis" (imitation). Genette defined these relations as the problem of "distance" between a narrator and a story or the information to be narrated. From the viewpoint, the state of diegesis means longer distance, and the one of mimesis means shorter distance. In the state of longer distance, we can find various characteristics or narrative techniques; the compression of information in a story, the emphasis of a narrator's existence or the exposure of a narrator itself, the insertion into discourse of the narration by a narrator itself, the usage of indirect speech in language expression, and so on. On the other hand, in the state of shorter distance too, we can find many discourse methods; descriptive and dramatic narration, excessively detailed description of events and objects, the usage of internal monologue, the usage of direct speech in language expression, and so on. Genette also describes the law of "the degree of a narrator's existence in a story + (descriptive quantity about the) story = equal".

Genette's proposition is a very conceptual idea, but we try to give the operational and precise definition in a distance's computational system [8]. Input data into the system are a sequence of events, and for each event, a user inputs the value of distance from 0 to 10. The system interprets that if the value is nearer 0, the degree of mimesis is higher, and if it is nearer 10, the degree of

diegesis is higher. Based on these data, the system processes next four kinds of operations.

At first, we prepare the mechanism renews an events sequence using a story tree. All leaves and nodes in the tree correspond to events, and higher level's nodes mean more abstract events, and lower level's ones mean more concrete events. A sequence of events as input data correspond to node(s) or leaf (leaves) in the story tree that the system preliminary holds. However, child nodes under a same parent are all included in the sequence of events or is not all included in the sequence. When the values of distance in events with a same parent are sufficiently high, the system removes all child events from the sequence of events, and moves up the hierarchy in the tree by one step, and then substitutes the child events into the place where the parent event existed. On the contrary, if the values of distance of a parent event in a sequence of events are sufficiently low, the system removes the parent event from the sequence of events and replaces the same place by the child events. Going up in a hierarchy means more abstract processing, namely longer distance. On the contrary, going down means more concrete processing, namely shorter distance.

Second, the system uses the function of event slot removing. It removes slot or slots in each event in the sequence of events after above operation. The number of slot in an event to be removed is determined by the rate based on the value of given distance. For example, if the value of distance is "4", forty percent slots are removed. The selection of the slot(s) to be removed is (are) decided at random.

Above two tasks are operations of distance inside a story's content itself, but distance is closely related with the problem of narrator's appearance and intervention. First, the system transforms the value of distance of each event to correspondent percentage, and interprets it as the probability for generating narrator's events. Next, the system determines original narration by the narrator. After the generation of narrator's events was determined, the system uses a knowledge frames network for determining narrative objects that does not have any relations with the story itself. The knowledge frames network is a knowledge structure that has events and various concepts linked to them like characters, places, objects, human relations, etc. Using this, the system decides the number of links to search according to the value of distance, and selects the slots in the knowledge frame that it finally reached as the object of narrator's event. At the same time, the system generates the narrator's mental actions like opinion and imagination using specially prepared slot's values in the event frame.

2.4 Duration's techniques

Duration means relations between temporal passage in a story and the quantity of the description. It is the flow of time like "slow" and "speedy" in a narrative. Genette divided duration into four types techniques; pause, scene, summary, and ellipsis. But he did not show more detailed or concrete classification. We have considered about lower level's techniques of these duration's categories utilizing the analysis of a novel and other documents too.

In pause, temporal progress in a narrative stops, and the discourse time is longer than story time. We can think, for example, next techniques for really doing it; (1) Explanation: the conceptual narration about narrative elements such as character, event, object, and so on. (2) Digression: the narration not having the relation with the story to be narrated. (3) Description: the detailed and concrete narration about character, event, object, and so on.

In scene, discourse time is equal to story time. The expression of scene is made by a narrator or a character. The expression by a narrator has next techniques; (1) A scene's, so to speak, simultaneous description or a kind of live announce or reporting: the narration about things occurring currently. (1-1) The narration (simultaneous description or reporting) about the mental situation or state in a narrator itself or other character(s) in the current scene. (1-2) The narration about the external situation or state in character(s) and other object(s) in the current scene. (2) The utterance by a character or characters. (2-1) Dialogue means the conversation among some characters. (2-2) Monologue means the utterance doing by only one person. (3) Thought or thinking is the internal monologue. The themes or contents introduced by these techniques include opinion, anticipation, desire, daydream, question, and so on.

In summary, discourse time is shorter than story time. (1) Report is narrating the result and the passage in an event. (1-1) Report on an event which had finished means a kind of recall of a past finished event. (1-2) Report on an event which has finished means the narration about an event continuing from a point in the past to current time.

In ellipsis, there is no part of the narrative text corresponding to a story. On this category, Genette himself showed some sub categories; explicit ellipses, implicit ellipses, and hypothetical ellipsis.

3 A Discourse Process by Narrative Discourse Techniques

The extreme purpose of my research about the expansion of narrative discourse is blending its aspects into a consistent computational system and combining with other narrative phases, namely story, expression, social or marketing, literary works, and so on, to create a whole narrative generation system. We consider on a consistent narrative discourse process using an incomplete trace of narrative discourse process. Some narrative discourse techniques we explained above were implemented as experimental simulation systems, but they are partial attempts inside each aspect and here we think about how to compose a consistent narrative discourse simulation by showing the mutual transformation relation between input text and output text in each technique.

Following list is the first input text which is really generated by the simulation program based on the literary theory of Propp [3]. This is a simple skeleton of a

story, and aspects of discourse are omitted. Real output is conceptual representation, but in following example, we use conveniently simple Japanese and English natural languages representation.

男の子の父は家から森へ男の子のために出かけた. (A boy's father went to a forest from a house for the boy.)/男の子の母は男の子を家に監禁した. (The boy's mother confined the boy in the house.)/男の子は家から脱出した. (The boy escaped from the house.)/司祭はお姫様に金の鳥を質問した. (A priest asked a princess about a golden bird.)/お姫様は司祭に金の鳥を教えた. (The princess taught the golden bird to the priest.)/司祭は王様に乱暴した. (The priest used violence to a king.)/王様は金の鳥に反応した. (The king reacted to the golden bird.)/司祭は金の鳥を食べた. (The priest ate the golden bird.)/男の子は家の近くから森へ向かって出立した. (The boy left from near the house toward the forest.)/男の子は森で蛇と闘った. (The boy fought with a snake in the forest.)/男の子は蛇を殺害した. (The boy murdered the snake.)/死者は男の子を捕えた. (A dead caught the boy.)/男の子は死者を供養した. (The boy held a memorial service for the dead)/死者は男の子に鋼の肉体を贈った. (The deadpresented the boy with a body of steel.)/死者は男の子を森から王国へ案内した. (The dead showed the boy to the kingdom from the forest.)/男の子は王国で司祭と闘った. (The boy fought with the priest in the kingdom.)/男の子は司祭に勝った. (The boy won the priest.)/男の子は王様に金の鳥を譲渡した. (The boy transferred the golden bird to the king.)/司祭は自殺した. (The priest committed suicide.)/男の子はお姫様と結婚した. (The boy married the princess.)

3.1 Temporal order's processing

The mechanism for temporal order's transformation first accepts the above sequence of events and transforms it into a different sequence of events using basically structural discourse techniques of time and operational discourse techniques of time. The system decides the kinds of structural discourse techniques, event(s) to be inserted and the place(s), and other necessary information, and changes the narrative structure considering some constraints. And the system applies operational discourse techniques to the transformed part. At last, the system outputs a sequence of symbolic description. Next text is an example of the transformation, and in it, the order events appear is changed, and indications for modifying it like "dream" are used.

お姫様は「男の子が男の子が蛇を殺害した夢を見る気がする」とつぶやいた. (A princess mumbled "I feel that a boy will dream that the boy murdered a snake.")/男の子の母は男の子を家に監禁した. (The boy's mother confined the boy in the house.)/熊は「男の子は死者を供養することになるぞ」とつぶやいた. (A bear mumbled "The boy will hold a memorial service for a dead.")/男の子は家から脱出した. (The boy escaped from the house.)/司祭はお姫様に金の鳥を質問した. (A priest asked the princess a golden bird.)/男の子は熊に「私は男の子が蛇を殺害した夢を見る予定だ」と話した. (The boy talked "I will dream the boy murdered the snake" to the bear.)/司祭は王様に乱暴した. (The priest used violence to the king)/男の子は男の子が蛇を殺害した夢を見た. (The boy dreamed the boy murdered the snake.)/王様は金の鳥に反応した. (The king reacted to the golden bird.)/司祭は金の鳥を食べた. (The priest ate the golden bird.)/男の子は家の近くから森へ向かって出立した. (The boy left from near the house toward a forest.)/男の子は森で蛇と闘った. (The boy fought with the snake in the forest.)/死者は男の子を捕えた. (A dead caught the boy.)/男の子は死者を供養した. (The boy held a memorial service for the dead.)/司祭はなぜか男の子が司祭に勝つ夢を見た. (The priest dreamed the boy won the priest somehow.)/死者は男の子に鋼の肉体を贈った. (The dead presented the boy with the body of steel.)/死者は男の子を森から王国へ案内した. (The dead showed the boy to the kingdom from the forest.)/男の子は王国で司祭と闘った. (The boy fought with the priest in the kingdom.)/男の子は蛇を殺害した. (The boy murdered the snake.)/男の子は司祭に勝った. (The boy won the priest.)/侍女たちは「男の子は蛇を殺害したらしい」と噂し合った. (The lady's maids mutually rumored "the boy seemed to murder the snake".)/お姫様は司祭に金の鳥を教えた. (The princess taught the golden bird to the priest.)/男の子は王様に金の鳥を譲渡した. (The boy transferred the golden bird to the king.)/男の子は家から脱出した. (The boy escaped from the house)/司祭は自殺した. (The priest committed suicide.)

3.2 Perspective's processing

Zero perspective can narrate all narrative information included in a story, and the system simply outputs the information inputted into the story as it is. Internal fixed perspective, first, outputs three kinds of information; information about action(s) in event(s) that a particular person who functions as the center of the perspective (focalized person) makes, external information described in the event(s), and internal information of the focalized person. Next, if this focalized person knows other's event, the system removes the internal information in the other's persona-frame because focalized person cannot perceive the other's internal information, and outputs the action's information and external information in the event(s). External perspective outputs only external action(s) which appear(s) outside and the information of external aspects among the narrative information included in the story data. And then, it removes mental action(s) (which does not appear outside, for example thinking and considering) of person and internal information.

For the use of internal fixed perspective, first, we give mental states into "boy" in above text. (In following description of examples, Japanese are omitted.)

The boy's mother confined the boy in the house. -> A boy's mother confined the *excited* boy in the house. / The boy escaped from the house. -> The *excited* boy

escaped from the house. / The boy talked "I will dream the boy murdered the snake" to the bear.) -> The *calm* boy talked "Anxious I will dream the *excited* boy murdered the snake" to the bear. / The boy dreamed the boy murdered the snake. -> The *anxious* boy dreamed the *excited* boy murdered the snake. / The boy left from near the house toward a forest. -> The *tense* boy left from near the house toward a forest. / The boy fought with the snake inthe forest. -> The *excited* boy fought with a snake in the forest. / A dead caught the boy. -> A dead caught the *scared* boy. / The boy held a memorial service for the dead. -> The *calm* boy held a memorial service for the dead. / The dead presented the boy with the body of steel. -> The dead presented the *calm* boy with a body of steel. / The dead showed the boy to the kingdom from the forest. -> The dead showed the *happy* boy to the kingdom from the forest. / The boy fought with the priest in the kingdom. -> The *desperate* boy fought with the priest in the kingdom. / The boy murdered the snake. -> The *excited* boy murdered the snake. / The boy won the priest. -> The *desperate* boy won the priest. / The boy transferred the golden bird to the king. -> The *flesh* boy transferred the golden bird to the king. / The boy escaped from the house. -> The *excited* boy escaped from the house.

Next text is an example that paralipsis is used to one of boy's events ("The *excited* boy murdered the snake.") processed by internal fixed perspective.

A princess mumbled "I feel that a boy will dream that the boy murdered a snake."/ A boy's mother confined the *excited* boy in the house. / A bear mumbled "The boy will hold a memorial service for a dead."/ The *excited* boy escaped from the house. / A priest asked the princess a golden bird. / The *calm* boy talked "Anxious I will dream the *excited* boy murdered the snake" to the bear. / The priest used violence to the king. / The *anxious* boy dreamed the *excited* boy murdered the snake. / The king reacted to the golden bird. / The priest ate the golden bird. / The *tense* boy left from near the house toward a forest. / The *excited* boy fought with a snake in the forest. / A dead caught the boy. / The *calm* boy held a memorial service for t he dead. / The priest dreamed the boy won the priest somehow. / The dead presented the *calm* boy with a body of steel. / The dead showed the *happy* boy to the kingdom from the forest. / The *desperate* boy fought with the priest in the kingdom. / *The boy murdered t he snake.* / The *desperate* boy won the priest. / The lady' s maids mutually rumored "the boy seemed to murder the snake" . / The princess taught the golden bird to the priest. / The *flesh* boy transferred the golden bird to the king. / The *excited* boy escaped from the house. / The priest committed suicide.

3.3 Distance's processing

We use distance operation to only a part in previous text, and show the concept of distance. First, an event is transformed into more concrete form using the moving of a story tree for realizing nearer distance. For example, we hypothesize that there is a small story tree of *"murder"*; *(murder (invite sleeping-pil-in-a-cup-of-wine sleep cut die))*. In this case, an event in the original text "The boy murdered the snake" is transformed into a more detailed sequence of events; *"The boy invites the snake. / The boy gave sleeping pil in a cup of wine. / The snake slept. / The boy cuts the snake's body. / The snake died."*.

Next, a narrator's event, *"Now my mother came home. / We wait continuing narration until my mother come to this room. / My mother entered into my room. / I start to narrate."*, is inserted into the point between, for example, "The boy's mother confined the boy in the house." and "A bear mumbled 'the boy will hold a memorial service for a dead."" for realizing very far distance.

3.4 Duration's processing

In this text, the figure of the princess in first event is described at little detailed way, and while this description is working, temporal progress of the event is paused.

A princess mumbled "I feel that a boy will dream that the boy murdered a snake."/ *The princess has blue eyes. / The princess has black hair. / The princess is tall. / The princess is wearing a white dress.* / A boy's mother confined the excited boy in the house. / Now my mother came home. / We wait continuing narration until my mother come to this room. / My mother entered into my room. / I start to narrate. / A bear mumbled "The boy will hold a memorial service for a dead."/ The excited boy escaped from the house. / A priest asked the princess a golden bird. / The calm boy talked "Anxious I will dream the excited boy murdered the snake" to the bear. / The priest used violence to the king. / The anxious boy dreamed the excited boy murdered the snake. / The king reacted to the golden bird. / The priest ate the golden bird. / The tense boy left from near the house toward a forest. / The excited boy fought with a snake in the forest. / A dead caught the boy. / The calm boy held a memorial service for the dead. / The priest dreamed the boy won the priest somehow. / The dead presented the calm boy with a body of steel. / The dead showed the happy boy to the kingdom from the forest. / The desperate boy fought with the priest in the kingdom. / The boy invites the snake. / The boy gave sleeping pil in a cup of wine. / The snake slept. / The boy cuts the snake' s body. / The snake died. / The desperate boy won the priest. / The lady' s maids mutually rumored "the boy seemed to murder the snake" . / The princess taught the golden bird to the priest. / The flesh boy transferred the golden bird to the king. / The excited boy escaped from the house. / The priest committed suicide.

4 Conclusions

In this paper, we treated the aspect of narrative discourse which is one of the most important and complex element in narrative generation, and tried to

show an approach toward a consistent computer simulation. We have many problems for the final goal, and show some of them.

- The problem of data structure like input and output data: As we independently consider and decide the mechanism and specification about each discourse techniques until now, data structures (mainly, input & output data and internal knowledge representations) used in each technique are respectively different. Thus, we can not still make a consistent computer simulation of discourse transformation. But, as it is related to the problem of how to define or distinguish the range of each phase (especially, story and discourse) in a narrative generation process, is difficult to easily decide. In principle, we think that story phase generates a story world (the story in the wide sense) to include all information which have potential possibility to be narrated and a sequence of events along temporal order (the story in the narrow sense). In the case of non sequential novel like hypertext novel, the generation of the story in the narrow sense may not be necessary. Therefore, narrative discourse processing is limited to pure transformation processing of story information. On the contrary, for deciding the kind of information which should be generated in story phase, the analysis of narrative discourse mechanisms and their organizational integration are necessary.
- The order of processing: We can think the order for applying narrative discourse techniques to text. For example, there is the sequential order which discourse techniques are applied in a fixed order like above process (temporal order -> perspective -> distance -> duration). We can also think a kind of distributed discourse processing in which various discourse techniques are executed as occasion arises to a part of text.
- Levels of narrative discourse techniques: As discourse techniques classified by Genette's are comparatively higher level's rhetoric and their many techniques do not have sufficient or direct operational functions, we have to expand Genette's model as the set of lower level's operational functions. Through narrative analysis, we could acquire or define more concrete techniques. For example, there are many lower level's techniques such as "description", "explanation", "simultaneous reporting", and so on. Many techniques of them are general operations which can be commonly utilized under some Genette's techniques. From a viewpoint, Genette's typology may mean a kind of literary rhetoric and lower level's techniques which give more concrete operations to it may mean cognitive or linguistic procedures. Integrating both elements means both cognitive/ computational expansion of literary theories and literary expansion of cognitive science/ artificial intelligence (we have insisted on "expanded literary theory" [1]).
- The blending of technical knowledge and strategic knowledge: Narrative discourse techniques explained in this paper do not include strategic knowledge on how to use them like the timing to be executed. Such knowledge affects the aspect of narrative structure and the aspect of cognitive effect in readers such as interestingness and emotional response. We currently want to avoid including easily psychological element into literary research, but it is necessary and important to introduce the strategic knowledge related to narrative structure from the viewpoint of at least literary effect. We want to start from the consideration of ad hoc and fragmentary description about strategic narrative knowledge in [5].
- The integration into a narrative generation mechanism: This is related to first problem. Based on narratology, narrative generation process is divided into two phases of story and discourse, but we divides discourse into two aspects; discourse in the narrow sense and surface expression from the point of computational symbol processing. Therefore, we have to decide the range in each processing of story, discourse, and expression. Especially, we think that story also has two distinguished aspects; story world (story in the wide sense) and a sequence of events (story in the narrow sense). Story world is a knowledge structure in which all information occurred in the fictional or virtual world, and may mean a kind of chronicle, and story in the narrow sense is a selective arrangement of in story world's information. It may include diverse types of knowledge and cover wide range, but we will be able to acquire various knowledge types through the functional analysis of each narrative discourse techniques.

References

[1] Ogata, T.: Expanded Literary Theory: Cognitive/Computational Expansion of Literary Theories and Narratology, *Proc. of 17th Congress of the International Association of Empirical Aesthetics*, 163-166, 2002.

[2] Propp, V. (Пропп, В. Я.): *Морфология сказки, Изд, 2е.* Москва: Наука, 1969.

[3] Ogata, T. & Terano, T.: Explanation-Based Narrative Generation Using Semiotic Theory, *Proc. of National Language Processing Pacific Rim Symposium 91*, 321-328, 1991.

[4] Hosaka,Y. & Ogata,T.: A Study of Story Generation and Representation - A Simulation of V.Propp Theory as a Starting Point -, *The 4th International Conference on Cognitive Science* (*Abstracts*, 64), 2003.

[5] Genette, G.: Discours du recit, essai de methode, *Figures III*, Paris: Seuil, 1972.

[6] Mukouyama, K., Shinohara, K., Kanai, A. & Ogata, T.: Rhetorical Analysis and Automatic Editing of the Film, *Proceedings of 17th Congress of the International Association of Empirical Aesthetics*, 571-574, 2002.

[7] Ueda, K. & Ogata, T.: Classification and Combination of Perspective in Narrative, AROB2004, 2004.

[8] Ogata, T. & Yamakage, S.: A Computational Mechanism of the "Distance" in Narrative: A Trial in the Expansion of Literary Theory, Unpublished (1-6), 2003.

Proc. of The Ninth Int. Symp. on Artificial Life and Robotics (AROB 9th '04)
Beppu, Oita, Japan, 28-30 January, 2004

Extraction of Meaningful Tables on the Internet*

Sung-Won Jung
Dept. of Computer Engineering
Pusan National University
Busan, Korea, 609-735
swjung@pusan.ac.kr

Jae-ho Kang
Dept. of Computer Engineering
Pusan National University
Busan, Korea, 609-735
jhkang@pusan.ac.kr

Kwang Ryel Ryu
Dept. of Computer Engineering
Pusan National University
Busan, Korea, 609-735
krryu@pusan.ac.kr

Hyuk-Chul Kwon
Dept. of Computer Engineering
Pusan National Univerisity
Busan, Korea, 609-735
hckwon@pusan.ac.kr

Abstract

The information retrieval system currently in use fails to consider the structural information of documents but uses extracted indexes from documents instead. Structural information such as the font face, font size, indentation, tables, and etc. demonstrate the author's meaning and is clearly the prime means of documentation. This paper pays special attention to tables because tables are commonly used within many documents to make the meanings clear, which are well recognized because web documents use tags for additional information. On the Internet, tables are used for the purpose of the knowledge structuring as well as design of documents. Thus, we are firstly interested in classifying tables into two types: meaningful tables and decorative tables. However, this is not easy because HTML does not separate presentation and structure. Therefore, this paper proposes a method of extracting meaningful tables using a modified k-means and compares it with other methods. The experiment results show that classifying on web documents is promising.

1 Introduction

The ultimate goal of an information retrieval system is to offer the most suitable information to its users. Performance is measured by its ability to find documents with useful information for the user. The related works have, therefore, focused mainly on the method of improving recall and precision. Special attention was given to web documents where an increasingly large number of users produce un-verified documents. An efficient retrieval method is required to improve the accuracy of such systems.

To make a high precision system, we must analyze the semantics of html documents. However, it is very difficult with present technology to apply semantics to an internet retrieval system. Another method, by which we grasp the author's intention, is to analyze the structural information of html documents.

* This work was supported by National Research Laboratory Program (Contract Number : M10203000028-02J0000-01510) of KISTEP.

This paper examines tables in several informational structures in html documents. The table form is more obvious than the plane text form, because we use a structured table in documents to convey our subject clearly. A performance improvement of an information retrieval system can be expected through the analysis of the tables because it is easy to find the tables and easy to extract the meanings in the web documents. This method can also be applied on the ranking models of current information retrieval systems like the vector space model, the P-norm model [1,2,3,4,5] etc.

2 The Challenge

Current information retrieval systems rank related documents based on similarities between documents and the users' query [1,2]. Such systems place great importance on index words and have the following limitations.

Firstly, current systems do not draw the meaning from html documents. To retrieve information more accurately, semantics of html documents should be considered. These information retrieval systems measure the similarity between html documents and the users query based on the 'term frequency' and 'document frequency'. Furthermore, these systems accept the assumption that documents are related to the index word in proportion to the 'term frequency' of the documents. However, this assumption is only of limited validity.

Secondly, current systems do not distinguish relatively weighted indexes or keywords from general indexes in html documents. Because there can be more important keywords, even in the document, people tend to remember the important part when they read a certain document. If all keywords in a document are given the same weight, it is difficult to retrieve exactly what the users want. The systems consider the 'term frequency' and 'document frequency' of the indexes by an alternative method but it does not increase the matching accuracy sufficiently enough to satisfy all users.

Lastly, the current systems do not reflect structural information in html documents. We can grasp the writer's intention to some degree if we consider the structural information of the documents. As the writer uses titles for each paragraph, indentations and tabular forms to convey his intention clearly, they have significance in a document. These information retrieval systems, however,

ignore the structural information when they extract index words from these documents.
In addressing these limitations the devised system aims to analyze and process tabular forms. For an author writing a document it may be more effective, therefore, to represent the document with a tabular form rather than to describe it in the usual methods. Related works [9, 10, 11, 12, 13] that study the extraction of table information handle a special tabular form, and extract information using abstraction rules. Similarities exist with this and our previous work [8]. The defect of such methods, however, is that it is difficult to apply them to various web document formats. In order to address this problem a method of extracting meaningful tables using decision trees and applying their results to an information retrieval system has been developed.

3 Implementation

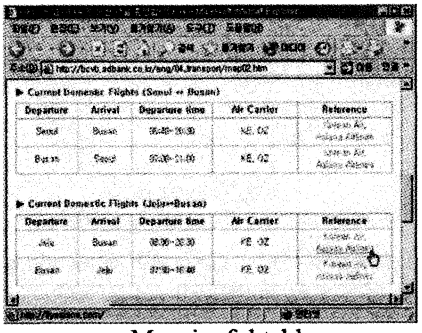
3.1 The classification of tables on the Internet

Various kinds of tables exist on the Internet. In general, a table is a printed or written collection of figures, facts, or information, arranged in orderly rows across and down the page which conveys an author's meaning more clearly. The <table> tag is supported in HTML for use on the web in a table. However, tables on the web are used not only for the purpose of the original goal but also to make the information clearer by lining it up. The task of deciding which tables are meaningful on the web is necessary before the information in a table can be extracted. To achieve this, the tables on the web have been classified into two types:
First, it is important to know which tables can be processed by the information retrieval system from amongst the meaningful tables on the web to extract the required table information. Figure 1.a is a web page from the Busan Convention and Visitor Bureau [7]. In this paper this kind of table is called the 'meaningful table', which can be defined as having the following criteria:

- The index of the table is located in the first row and the first column.
- The contents of a cell extracted by a combination of rows and columns have specific information.
- The term and document frequency cannot represent a relation between row index, column index and the content of a cell in a table.

Besides the tables having the above characteristics, we can often usually see a web page like the one in Figure 1.b. The <table> tags are used all over the HTML source of this document. However, the tables of Figure 1.b are not illustrating the owners' meaning but are offering a well-arranged display to the users. This kind of table is defined as the 'decorative table' and the following criteria:

- The table does not have table indexes in the first row and the first column.
- The table has no repeatability of cell form, and is not structural.
- The contents of the table have complex contents such as long sentence, image, etc.



a. Meaningful table

b. decorative table

Fig. 1. The types of table on the Internet

3.2 Institution of table features

It is unusual to encounter a document which includes meaningful tables on the web. Nevertheless, the meaningful tables should be extracted from amongst the others for effective information retrieval; however, it is impossible to extract them manually due to the large number of internet documents. Therefore, the tables should be defined automatically, to identify if they are meaningful or not on the basis of the characteristics which were obtained previously by analyzing meaningful tables in the training data. For example, several characteristics could be acquired from Figure 1: presence of pictures, presence of background colors and the table size.
The whole features used in the devised system are as follows:

1. The presence of <caption> tag
2. The presence of <th> tag
3. The presence of <thead> tag
4. The classification of the cells on the first rows
5. The difference of the background color option between the row with the index and the next row, or the difference of the background color option between the column with the index and next column

6. The difference of the font color option between the row with the index and next, or the difference of the font color option between column with index and next column
7. The difference of the font face between the row with the index and next row, or the difference of the font face between column with the index and the next column
8. The presence of the border option
9. The ratio of empty cells
10. The ratio of cells including <a href> tag
12. The ratio of cells consisting of text only
13. The presence of rows or columns consisting of numeric data only
14. The ratio of cells consisting of only symbols
15. The presence of other tables in the table
16. The table size
17. The number of sentences with more than 40 characters in a cell
18. The type of table shape
19. The difference of the font size between the row with index and next row, or the difference of the font size between the column with the index and the next column
20. The index word type
21. The number of meaningful index column

3.3 Application of Proposed Learning Algorithm

We convert the features, which have been extracted from a table, to vector of 21 dimensions. These vectors are input data of machine learning algorithms. Our previous work[14] applies this data to ID3 algorism using information gain. The limitation of our previous method is that it has nominal values as results. Thus, we were unable know which degree of ambiguity a table shows. Therefore, modified k-means algorithm is used in this work because we need continuous value. Compared with decision tree algorithm, our new algorithm has following advantages:

1. This method can be used to determine the level of ambiguity existing in a table.
2. We can get representable types of tables.
3. This method can apply to the extraction table information.
4. This method can easily combine other information retrieval methods.

Figure 2 illustrates our proposed method. Let's observe the process step by step. Firstly, each feature value is changed into its frequency ratio by numeric formula of step 1. This ratio means classification power of each feature. For the test we change our extract data from tables into a modified one with continues value. This modified learning data set, is foundation data for proposed method. Secondly, data set is divided into two subsets, meaningful tables and decorative tables. Each subset is clustered by k-means algorithm and we get centroids of clusters which are representable type of each subset.

Step 1. Making Learning Data Set

$$S=\{V_1,V_2,\cdots,V_n\} \quad V_i=\{v_1,v_2,\cdots,v_{21}\} \quad v_i=\{a_{i,1},a_{i,2},\cdots,a_{i,m}\}$$
$$\Rightarrow\ p(a_{i,j})=\frac{S(a_{i,j})}{N}\ \Rightarrow$$
$$\hat{S}=\{\hat{V}_1,\hat{V}_2,\cdots,\hat{V}_n\} \quad \hat{V}_i=\{w_1\hat{v}_1,w_2\hat{v}_2,\cdots,w_{21}\hat{v}_{21}\} \quad \hat{v}_i=\{p(a_{i,1}),p(a_{i,2}),\cdots,p(a_{i,m})\}$$

S : real data set

$\hat{S}$: modified real data set

V_i : a real datum with vecotr of 21 demenstion, i.e. a table

$\hat{V}_i$: a modified real datum

v_i : attribute of vector

$\hat{v}_i$: modifined attribute of vector

w_i : coefficent of modifined attribute

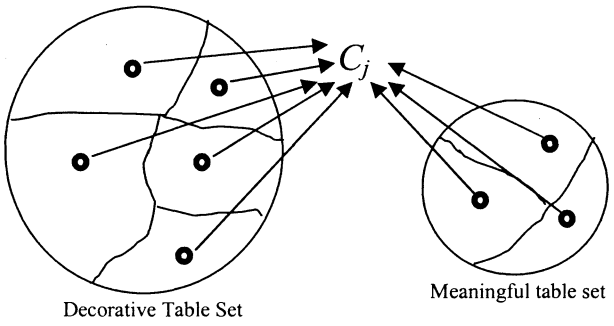
$a_{i,j}$: value of i attribute

$p(a_{i,j})$: frequent ratio of $a_{i,j}$

$S(a_{i,j})$: Number of $a_{i,j}$ in learing data set

N : Number of learing data

Step 2. Get centroids of representable type



Step 3. Get coefficient for classification

$$M(\hat{V}_i)=\begin{cases}1:\hat{V}_i \text{ is meaningful table}\\ 0:\hat{V}_i \text{ is decorative table}\end{cases}$$

$$E(\hat{V}_i)=\begin{cases}1:\underset{C_j}{\operatorname{argmax}}\,D(\hat{V}_i,C_j),M(\hat{V}_i)=M(C_j)\\ 0:\underset{C_j}{\operatorname{argmax}}\,D(\hat{V}_i,C_j),M(\hat{V}_i)\neq M(C_j)\end{cases}$$

$$\arg\max R(\hat{S})=\arg\max_{w_1,w_2,\cdots,w_{21}}\sum_i E(\hat{V}_i)$$

Fig. 2 Proposed Learning Algorithm

3.4 The System implementation

Figure 3 is a system implementation procedure. The whole procedure of the task is summarized as follows. Firstly, the sample data to be analyzed is collected and is then classified by tables and the others by hand. We investigate 100,000 documents to establish the sample data for this system. Secondly, we implement the system which extracts the features of each table automatically after the table information is extracted from the parsing of HTML documents. The sample data is generated from the merging of the former and latter results. Using this sample data, our system implements our proposed

method and creates a table information dictionary. Finally, the system uses this dictionary for the information retrieval system.

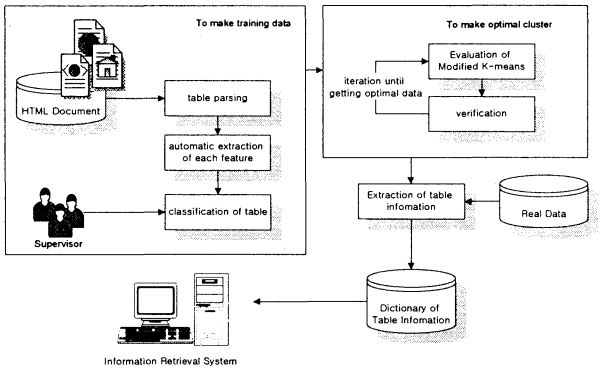


Fig. 3 The process of the system implementation

4 Experimental Result

We implemented the system with an Intel Pentium IV that has a 2GHz CPU and 256 MBytes main memory. Sample data was extracted in about 100,000 internet documents and test data was chosen from about 5 millions internet documents that were collected in Artificial Intelligence Laboratory in Pusan National University.

Table 1 shows error rates according to variation of cluster number. The cluster number coincides with the number of representable type for each class. From Table 3, we set MN for 2 and DN for 3, because we judge one class with only one cluster to overfit.

Table 1. Error Rate for Number of Cluster

No. of Table : 9769		No. of Decorative Table Cluster(DN)				
		1	2	3	4	5
No. of Meaningful Table Cluster(MN)	1	247	106	60	40	48
	2	199	92	53	54	54
	3	131	113	98	71	54
	4	111	119	96	99	86

Table 2 is a result from the application of the modified k-means to test data. The data is 5,000 web documents collected on the Internet. For the results a recall of 84.58% and a precision of 75.52% were obtained

Table 2. The application of the modified k-means

Number of tables	Decision Tree		Modified k-means	
	Recall	Precision	Recall	Precision
4283	81.54	55.21	86.92	79.58
8020	52.91	55.56	83.60	72.15
6341	67.47	39.72	78.31	62.50
9108	86.11	63.70	85.65	83.33
9769	95.83	38.33	91.67	68.75
37521	73.36	54.45	84.58	75.52

5 Conclusions and Future Work

The devised information retrieval system requires making estimates to show correct information to the user. For the system to extract information from a table, it must classify whether it is a meaningful table or a decorative table. If a meaningful table is extracted, it is easy to process the information in the table to other data forms. For example, when a user inputs a query such as the "Temperature of Busan", the system should output the correct result using abstracted table information. This work can process data of a general information retrieval system or a specific domain that has many tables. The work needs a more accurate algorithm for classifying and learning a table thereafter.

Future work will seek to identify the correlation between documents and tables. The rationale behind this work will research the link between tables and documents i.e. if a table can supply the contents of a document, how relevant tables can affect the importance of a document.

References

1. M. Kobayashi, K. Takeda, "Information Retrieval on the Web", ACM Computing Surveys, pp. 144-173, 2000
2. G. Salton, M. J. McGill, "Introduction to Modern Information Retrieval", McGraw-Hill, 1983
3. E. A. Fox, "Extending the Boolean and Vector Space Models of Information Retrieval with P-norm Queries and Multiple Concept Types", Dissertation Cornell University, 1983
4. M. E. Smith, "Aspects of the P-norm Model of Information Retrieval : Syntactic Query Generation, Efficiency, and Theoretical Properties", Dissertation Cornell University, 1990
5. G. Salton, E. A. Fox, H. Wu "Extended Boolean Information Retrieval", ncstrl.cornel, pp. 82-511, 1982
6. T. M. Mitchell, "Machine Learning", McGraw-Hill, pp. 53-79, 1997
7. http://www.busancvb.org/eng/home.html
8. S.W. Jung, K. H. Sung, T.W. Park, H.C. Kwon, "Effective Retrieval of Information in Tables on the Internet" IEA/AIE June pp. 493-501, 2002
9. J. Hammer, H. Garcia-Molina, J. Cho, R. Aranha, and A. Crespo, "Extracting Semistructured Information from the Web", SIGMOD Record, 26(2), pp. 18-25, 1997
10. Y. Huang, G.Z. Qi, F.Y. Zhang "Constructing Semistructed information extractor from the Web document", Journal of Software 11(1) pp. 73-75, 2000
11. N. Ashish, C. Knoblock "Wrapper Generation for Semi-structed Internet Sources", SIGMOD Record, 26(4) pp. 8-15, 1997
12. D. Smith, M. Lopez, "Information Extraction for Semi-structed Documents, In Proceedings of the Workshop on Management of Semistructed Data", Conjunction with PODS/SIGMOD, Tucson, AZ, USA, May, 12, 1997
13. G. Ning, W. Guowen, W. Xiaoyuan, S. Baile, "Extracting Web table information in cooperative learning activites based on abstract semantic model, Computer Supported Cooperative Work in Design, The Sixth International Conference on 2001 (2001) 492-49