

A Study on Extracting Ideas from Documents and Webpages in the Field of Idea Mining*

아이디어 마이닝 분야에서 문헌과 웹페이지의 아이디어 발췌에 대한 연구

Tae-Young Lee(이태영)**

ABSTRACT

The ideas and quasi-ideas useful for human's creation were drawn out from documents and webpages with extraction methods used in idea mining, opinion mining, and topic signal mining. The extraction methods comprised (1) decisive cue phrases, (2) cue figures and sounds, (3) contextual signals, and (4) discourse segmentations. They tested on the idea samples, such as thoughts, plans, opinions, writings, figures, sounds, and formulas. Methods (1), (3), and (4) received largely positive evaluation, judging the efficiency of 4 methods by F measure, a mixture of recall and precision ratio. In particular, decisive cue phrase method was effective to search idea and contextual signal method was effective to detect quasi-idea.

초 록

일반적인 문헌/문서나 웹페이지에서 창조에 도움이 되는 아이디어와 준아이디어를 색출하기 위하여 아이디어 마이닝 기법을 적용하였다. 아이디어 마이닝과 의견 마이닝 및 논제 신호 마이닝에서 사용하는 발췌 기법으로 웹 페이지, 문헌, 문서 등에 포함되어 있는 아이디어를 발췌하였다. 발췌 기법을 (1) 결정적 단서 어구, (2) 단서 멀티미디어, (3) 문맥 신호, 및 (4) 담화 구절 방법으로 정리하여 7가지 아이디어 유형 -사상, 계획, 의견, 글, 그림, 소리, 공식 별로 실험하였다. 각 기법들의 효율성은 재현율과 정확률을 혼합한 F 측정값으로 판단하였고 (1), (3), (4) 방법은 대체로 긍정적인 평가를 얻었다. 특히, 결정적 단서 어구는 아이디어 적출에 문맥 신호는 준아이디어 추출에 효과적인 것으로 판단되었다.

Keywords: idea mining, decisive cue phrase, cue multimedia, contextual signal, discourse segmentation, extraction method

아이디어 마이닝, 단서 어구, 단서 멀티미디어, 문맥 신호, 담화 구절, 발췌 기법

* 본 연구는 전북대학교 2010년도 후반기 인문·사회계열 교수 연구기반 조성비로 이루어졌음.

** Professor in library & information science dept. of chonbuk national university(taehyun@jbnu.ac.kr)

■ 논문접수일자: 2011년 12월 30일 ■ 최초심사일자: 2012년 1월 25일 ■ 게재확정일자: 2012년 1월 31일
■ 정보관리학회지, 29(1), 25-43, 2012. [<http://dx.doi.org/10.3743/KOSIM.2012.29.1.025>]

I. Introduction

The year before last, the government had announced the plan that fostered the creative company handled by one person as new growth engine in future and established tentatively named 'IBB (idea biz bank)' in order to collect, select, and evaluate the personal ideas on cyber space (Dong-A Daily News, 2009, p. 1). As shown the government plan, the promotion of software industries like IBB is important because more and more text streams are being generated in various forms, such as news streams, weblog articles, emails, instant messages, research paper archives, web forum discussion threads, and so forth (Wang, Zhang, Jin, & Shen, 2008). These texts have the potential to act as rich sources for raw inputs to market research and knowledge discovery. Since internet is a crucial driving force in today's world, these texts are rich pointers to the collective idea of the global population on almost every topic. Given the volume and growth rate of these sources, efficient mechanisms are required to aggregate, assimilate and interpret all the information with minimal human intervention (Dey & Haque, 2008).

Up to now, the identification of new ideas and inventions in unstructured texts is done manually (that means by humans) without the support of text mining. A new idea is often needed to discover unconventional approaches e.g. to create a technological breakthrough. However, a manual extraction of new ideas from these masses of texts is time consuming and costly. Therefore, it is useful to search for new problem solution ideas automatically or semi-automatically.

Therefore we will describe the theoretical background of the text mining approach to discover semiautomatically textual patterns representing ideas and inventions such as new thoughts, maxims, images, etc. in unstructured technological texts (Thorleuchter, 2008).

We have proposed a hybrid approach to mine ideas from noisy text data comprising blogs, on-line reviews, and customer feedbacks on products etc (Dey & Haque, 2008). In this paper. Idea mining from noisy text comprises the following sub-tasks with four extraction methods:

- (i) Firstly, a 'Decisive cue phrase' method commonly draws out the sentences or paragraphs, which contain cue words and phrases that notice a text stream entered is an idea. Actually, paragraphs which include sentences containing the word idea have the high probability to express the idea.
- (ii) Through 'Multimedia' method, pictures, diagrams, tables and mathematical formulas, which include ideas, are distinguished from the text stream entered. For instance, when a picture illustrates a person who shows happy expression, this system regards that as the idea that notices happiness. Formulas illuminating certain theories are accepted as a researcher's expression of the idea, recipes of foods are regarded to belong to cooks' idea and diagrams indicating inventions are comprehended as the idea of the invention.
- (iii) Through 'Contextual signal method', if some words - which are not decisive cue words, but frequently emerged in sentences- are

emerged in certain paragraphs, they are presumed as ideas with almost the same treatment as similar as ‘Corpus method’ which is generally used in ‘Automatic indexing’. At least, they will be regarded in the preparation step in which ‘ideas’ can be made up.

- (iv) Through ‘Discourse segmentation method’, ideas can be found in sentences by the same way of above (iii). There are ideas delivered by conversations as well as impressive contents which make people get an idea among conversations.

The rest of the paper is organized as follows. Section 2 presents related works and theoretical foundation, section 3 presents the types, characteristics, and examples of ideas treated in this study. Section 4 defines the extraction methods, section 5 presents experimental results and finally section 6 presents conclusions.

2. Related Work

Not much work has been done yet for idea mining but opinion mining is becoming popular research area in text mining community for its wide applicability (Dey & Haque, 2008), which is accepted as a type of a idea by various dictionaries. And also, there are the subject areas; ‘Topic Extracting’ (Buitelaar & Eigner, 2008; Chung & Kim, 2008; Wang, Zhang, Jin, & Shen, 2008) and ‘Document Summarizing’ (Brandow, Mite, & Rau, 1995; Edmundson, 1998;

Hovy & Lin, 1999; Kupiec, Pedersen, & Chen, 1995; Lee, 2005; Teufel & Moens, 1999), applying the similar treatment technology used in ‘Opinion Mining’ (Al-Halimi, 2003; Dave, Lawrence, & Pennock, 2004; Dey & Haque, 2008; Lee & Chung, 2009; Liu, 2009; Pang & Lee, 2008). The opinion mining, topic extracting, and document summarizing have employed similar methods, which prepare the cue word list, to decide on whether the word is important or not and to measure similarity between each other context of text.

In topic extracting or document summarizing, the basis of the method used corpus starts from an assumption that the information of sentence structures or the semantic information of words and phrases can represent the importance of a sentence. Edmundson (1998) had suggested four methods for ‘Automatic Abstracting’. Among them, cues, title words, and locations were the methods with corpus; The Cue method judged sentences including cue words like “thus, consequently, eventually or therefore” as more important sentence than others. After the suggestion, although there were some methodological modifications, Kupiec, Pedersen, and Chen (1995, p. 56), Myaeng and Chang (1999) and Hovy and Lin (1999) also referred to the corpus method.

Brandow, Mite, and Rau (1995) had selected ‘Signature’ word which passed certain thresholds after the calculation of words weights by applying the formula ‘tf/idf’, they also configured headline words as the signature words. And then they judged whether or not the sentences were considerable with the weights of the signature words located in one

sentence adding. Except these signature words, there were ‘Location in documents’, ‘Existences of anaphora’, ‘Length of extracts’, ‘Sorts of extracts’ and so on, which were regarded as conditions included into ‘Text (document) Summarization’ by Brandow (1995) and others. They also added sentences into extracts until the size of extract is reached to the standard size of the general summaries, and their allowable error range was approximately 10 words.

After Meadow, Boyce, and Kraft (2000) had configured factors like “n (the number of words in sentence)”, “k (keyword tokens in sentence)” and “kq (keyword types in sentence)” for distinguishing sentences, they distinguished significant and meaningless sentences by proportional ratio with relations among the factors such as “① k/n (the number of keyword tokens/the total number of words)”, “② kq/n (the number of keyword types/the total number of words)” and “d (the distance among keywords)” configuring.

Kupiec, Pedersen, and Chen (1995) proposed a formula, which measured the probability that a certain sentence was included in a summarized text, like below <Equation 2-1>. This equation was based on Bayesian’s classification (Mani, 2001, p. 60), and the length of sentences, cues, subject materials, idioms, topic-words and capital words were commonly contained in its Feature.

$$P(s \in E | F_1, \dots, F_n) = \frac{\prod_{i=1}^n P(F_i | s \in E) P(s \in E)}{\prod_{i=1}^n P(F_i)}$$

-- <Equation 2-1>

Where $P(s \in E | F_1, \dots, F_n)$: the probability that sentence s in original forms of texts will be included into an extract ‘E’ in the middle processing stage, $P(s \in E)$: the proportion of the extraction (constant), $P(F_i | s \in E)$: the probability of the feature F_i which is emerged at a sentence in an extract, $P(F_i)$: the probability of the feature F_i which is emerged at the corpus of the original texts, n : the number of features, F_i : the feature in order i .

And Al-Halimi (2003) suggested below <Equation 2-2> applied term frequency.

$$M^{TF}(w, t) = \log \frac{TF(w, t) * \|db\|}{TF(w, db) * \|t\|}$$

$$= \log \frac{TF(w, t)}{TF(w, db)} + \log \frac{\|db\|}{\|t\|}$$

-- <Equation 2-2>

Where

- TF (w, t) is the number of occurrences of w in topic t documents.
- TF (w, db) is the number of occurrences of w in the database db.
- ||t|| and ||db|| are the number of terms in t and db respectively.

<Equation 2-2> assumes that strongly relevant words are those that occur more frequently in topic t than in the whole database. On the other hand, a variety of methods based on basis of discourse · knowledge have been presented. Among them, Barzilay and Elhadad (1997) used the information of ‘Lexical chaining’ based on cohesion and coherence of vocabularies as the means for extracting important

sentences in order to summarize the text. Adapting ‘Corpus method’ and ‘Discourse structural method’, Teufel and Moens (1999) approached ‘Automatic extracting system’ with the comprehensive view.

Our work is closely related to Thorleuchter (2008) and Thorleuchter, Van den Poel, and Prinzie's (2009) work on idea mining. This approach introduces idea mining as process of extracting new and useful ideas from unstructured text. They use an idea definition from technique philosophy and focus on ideas that can be used to solve technological problems. To realize the processing, they use methods from text mining and text classification (tokenization, term filtering methods, Euclidean distance measure etc.) and combine them with a new heuristic measure for mining ideas.

3. Characteristics of Idea

3.1 Definitions and Types

An idea is, in the most narrow sense, just whatever is before the mind when one thinks. Very often, ideas are construed as representational images; i.e. images of some object. In other contexts, ideas are taken to be concepts, although abstract concepts do not necessarily appear as images. Thus, an idea can be a something, such as (1) a thought, notion, or conception, that potentially or actually exists in the mind as a product of mental activity. As all cognition is by ideas, it is obvious that the question of the validity of our ideas in this broad sense is that of

the truth of our knowledge as a whole (Wikipedia, n.d.).

We also find various definitions from the existing dictionaries that an idea is (2) the transcript, image, or picture of a visible object, that is formed by the mind (Webster online dictionary, n.d.). And (3) an opinion or belief (businessdictionary.com, n.d.), (4) a plan, scheme, or method (the free dictionary, n.d.), and (5) music a theme or figure (yourdictionary.com, n.d.) are picked up as the idea and elaborated it. Also, (6) a fiction object created by the imagination (Webster online dictionary, n.d.), and (7) estimate, estimation, approximation-an approximate calculation of quantity or degree or worth are designated to be a type of idea (definitions.net, n.d.).

As shown in the <Table 1>, an idea is a complex sense, the meaning of which can comprises seven categories like a “thought/conception, plan/design, opinion/belief, writing (fiction object) created by imagination, theme figure (picture), theme sound (music), and formula for estimation”.

Each of these idea types are now used in the right place of the documents but we need to know their characteristics and classes of which each idea type belong to, more detailedly to identify correctly in the field of web. Drawing on the above <Table 1>, idea classes were classified into the seven categories.

- (1) A thought class has a thought along with a concept, notion, and so on.
- (2) A plan class comprises a plan together with a design, scheme, syllabus, and so on.

<Table 1> The types of a idea

Dictionary \ Type	Thought/ Conception	Plan/ Design	Opinion/ Belief	Writing (Fiction/ Fancy Notion)	Figure	Sound	Formular (Estimate, Law, Recipe Etc.)
Wikipedia the free encyclopedia	○				○	○	
http://examples.yourdictionary.com/idea	○	○	○	○	○	○	
http://www.thefreedictionary.com/idea	○	○	○	○	○	○	
http://www.definitions.net/definition/idea (princeton's wordnet)	○	○	○		○	○	○
http://www.definitions.net/definition/idea (Webster dictionary)	○	○	○	○	○	○	
http://www.businessdictionary.com/definition/idea.html	○		○	○			

- (3) A opinion class contains a opinion in conjunction with a belief, view, review, prescription, conclusion, suggestion, and so on.
- (4) A writing has a writing in company with a maxim, humors, short story (conte), patent, article, and so on.
- (5) A figure class comprises a drawing, picture, image, icon, and so on.
- (6) A sound class includes a music a theme, song, music note, speech, conversation, video and so on.
- (7) A formular class has a two subclass, one is mathematical formular, the other is text formular. A mathematical law, theory, model, module, axiom, equation, etc. are belonged

to a mathematical formular and a textual law, rule, theory, recipe etc. are embraced in a textual formular.

3.2 Idea Examples

From web-sites and documents, actual examples were collected in terms of seven classes of ideas which had been categorized in the section 3.1. The extracting principle was constructed to perform the extraction according to the methodology used in the experimental evaluation which will be revealed in chapter 5. The collected examples are indicated in the below <Table 2>.

<Table 2> Examples of classes of idea

Class	Example	Remark
Thought	* Everyone has the right to freedom of thought, conscience and religion: this right includes freedom to change his religion or belief, and freedom, either alone or in community with others and in public or private, to manifest his religion or belief in teaching, practice, worship and observance.	thought, concept, notion, etc.

Class	Example	Remark
Plan	<p>* The structure of the facts book will be designed to match the specific needs of the organization, but one simple format – suggested by Malcolm McDonald – may be applicable in many cases. This splits the material into three groups:</p> <p>1. Review of the marketing environment. 2. ... 3. Review of the marketing system, ...current marketing objectives and strategies.</p> <p>Portfolio planning. In addition, the coordinated planning of the individual products and services can contribute towards the balanced portfolio.</p> <p>80:20 rule. To achieve the maximum impact, the marketing plan must be clear, concise and simple. It needs to concentrate on the 20 percent of products or services, and on the 20 percent of customers, that will account for 80 percent of the volume and 80 percent of the profit.</p>	<p>plan, design, scheme, syllabuses, etc.</p>
Opinion	<p>* George Washington didn't smile because of his false teeth</p> <p>* For proponents of this view, concepts mediate between thought and language, on the one hand, and referents, on the other. An expression without a referent ("Pegasus") needn't lack a meaning, since it still has a sense. Similarly, the same referent can be associated with different expressions (e.g., "Eric Blair" and "George Orwell") because they convey different senses. Senses are more discriminating than referents. Each sense has a unique perspective on its referent—a unique mode of presentation. Differences in cognitive content trace back to differences in modes of presentation.</p>	<p>opinion, belief, view, review, prescription, conclusion, suggestion, etc.</p>
Writing	<p>* proverb : Never use the passive where you can use the active / Say an old thing in a new way or a new thing in an old way.</p> <p>* short story : "That's easy. 'Mother's Day Sunday May 11. Remember your mother and she'll remember you," "Fine, you win," Sam said with disgust. Jan still had her eyes closed. "Come on, try again. Go ahead, ask me something else." "All right," Sam said. "What color socks am I wearing?" Jan thought a moment. "That's not really fair," she said. "I never saw your socks." But Jan didn't open her eyes. "You're wearing green pants, a green belt, and green shoes, so I'll bet your socks are green, too." "You're just too much, Jan!" "No, you're just too neat!"</p>	<p>maxim, humors, short story (conte), patent, article, etc.</p>
Figure		<p>drawing, photo- picture, image, icon, etc.</p>
Sound		<p>theme music, song, music note, speech, conversation, etc.</p>
Formular	<p>** mathematical formular: * Equation:</p> $\operatorname{argmax}_{g(t)} \sum_{m=1}^M \sum_{\mathbf{w}} \sum_{s=1}^T Q(\mathbf{w}, s) \sum_{\{d \in S_m : g(t)=s\}} c(\mathbf{w}, d)$ <p>** text formular: * Law: Homicide Act 1957, s. 2: Where a person kills or is party to the killing of another, he [sic] shall not be convicted of murder if he was suffering from such abnormality of mind (whether arising from a condition of arrested or retarded development of mind or any inherent causes or induced by disease or injury) as substantially impaired his mental responsibility for his acts and omissions in doing or being a party to the killing.</p>	<p>formular, law, rule, theory, recipe, etc.</p>

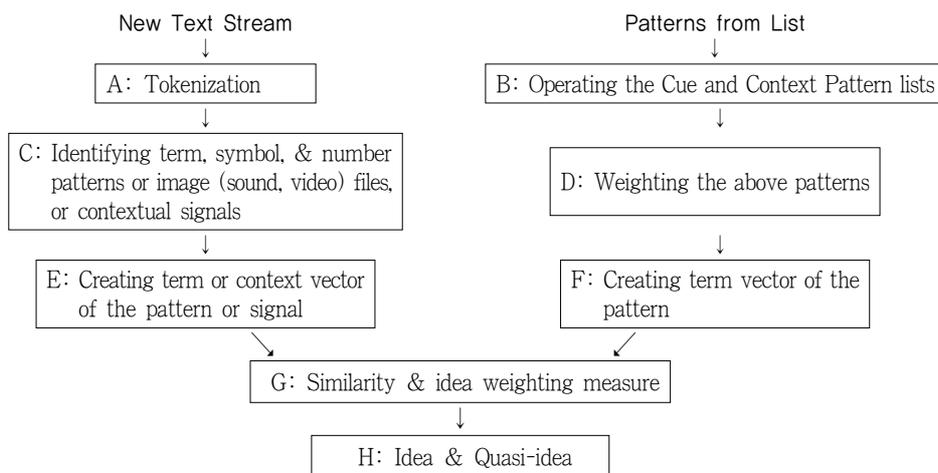
3.3 Quasi-idea

This study defines ‘Quasi-idea’ as a thought in the preparation step in which ideas can be made up and a thought being something missing to be an idea. ideas appear on text stream with the definite cue words and phrases such as “patent, law, proverb, trademark, this idea, in my opinion, as good idea etc.” or through the media such as “poem, picture, score, analects etc.”. On the other hand, quasi-ideas tend to appear with the signature words like “suggestion, opinion, creation etc.” while the paragraph contained quasi-idea is in high cohesive state. Pictures, musics, mathematical formulas, however, were classified into two categories, idea and no idea because they could be scarcely distinguished between quasi-idea and idea, while thoughts, plans, opinions, writings, textual formulas were classified into three categories, idea, quasi-idea, and no idea (differentiating between three categories was as-

signed to judging of graduate students).

4. Idea Mining Process

Several techniques for idea identification have been introduced in this literature, including methods based on cue phrases, cue figures like pictures and musics, contextual signals, discourse segmentations (Marcu, 1999). Our mining process will eventually contain modules that employ each of these methods. We introduce sets of domain specific terms which occur in the context of a term or a combination of terms. To date, modules for cue phrase, cue multimedia, contextual signal, discourse segmentation have been implemented. When it comes to the analyzing process (which is commonly applied) of the above three methods; cue phrase, contextual signals and discourse segmentation, it is stated the below <Fig. 1>



<Fig. 1> Common processing of our idea mining approach in different steps

A: With tokenization, texts are separated in terms and the term unit is word. The set of different terms in a text is reduced by using stop word filtering methods and stemming. For this, a general list of stop words is used as well as the well-known Porter stemming algorithm (Thorleuchter, Van den Poel, & Prinzie, 2009).

B: Adding and deleting are continuously repeated after the cue phrases and context patterns were listed.

C: Here, we show how to create these text patterns automatically, manually, and semi- automatically. Around each appearance of each term in the new text, we create a text pattern containing the selected term and all terms, which occur in the left and right context of the selected term. To reduce the number of text patterns, we only create text patterns around non-stop words and around terms that occur both in the new text and in the problem description (Thorleuchter, Van den Poel, & Prinzie, 2009).

D: Weights are assigned in each pattern, and they are updated through feedbacks.

E: For each text pattern from the new text, we create a term vector in vector space model. The size of the vector is defined by the number of different stemmed and stop word filtered terms in the new text. For text pattern encoding, we use binary term vectors that means a vector element is set to one if the corresponding unstemmed term is used in the text pattern and to zero if the term is not. We also build text patterns from the problem description and create term vectors as described above.

F: Each patterns are modified into the form of 'Binary term vector'.

G: We measured the similarity between E vector and F one, using the following formulas, <Definition 1, 2, 3-*, and 4> which is indicated in section 4.1-4.4 and was established by applying the equations in chapter 2. In this idea mining approach, we do identify synonyms and homonyms.

4.1 Decisive Cue Phrases

Certain cue phrases explicitly suggest, in automatic summarization, that what follows is in fact a summary or the main point of an article. Passages containing such cues should therefore be preferred for selection. Examples of summary cue phrases include: "in summary, to sum up, the point is, etc.". This method using cue phrases are also able to apply to identify an idea as well as to select summary. The phrases, "suggest idea, good idea, etc." give us the determinate evidences that the paragraphs included them as an idea.

In one experiment, we manually compiled a list of decisive cue phrases from a training corpus of paragraphs that themselves were texts of ideas. In this corpus, sentences containing phrases such as "new idea, good idea, presented the idea, and we conclude" fairly reliably reflected the major content of the paragraphs. This indicated to us the possibility of identifying a idea. <Fig. 2> contains an example, with sentences containing decisive cue phrases and cue words underlined.

Referring to the <Fig. 1>, these approaches, idea mining of decisive cue phrase method consists of three steps as like as follows and the same discipline like the <Definition 1> is applied here.

* paragraphs contain the idea representing words like idea(s), belief, opinion, notion, etc.,
 * this is a new idea, * in my opinion, * I believe, * we conclude, * create, creation, * design,
 * new, * axiom, * formular, * equation, * figure, * picture, * music

<Fig. 2> The decisive cue phrases and cue words leaded to ideas

- 1) Preparation of cue phrases list and term vectors of decisive cue phrases
- 2) Extraction of n-gram word sets from a new text and convert into term vectors of n-gram word sets
- 3) Comparison between term vectors of n-gram word sets and term vectors of decisive cue phrases and weighting the n-gram word sets

Definition 1. Let (a text) $CP = [cp_1, \dots, cp_\mu]$ be a list of decisive cue phrases (words) cp_i in order of appearance and let $\mu \in M$ be the number of phrases in CP and $i \in [1, \dots, \mu]$. Let $\gamma \in \Gamma$ be the number of words in cp_i and $n \in [1, \dots, \gamma]$. Let $TP = [tp_1, \dots, tp_\nu]$ be a set of n-gram text patterns (n-gram words) tp_j in a new text stream and let $\nu \in N$ be the number of patterns in TP and $j \in [1, \dots, \nu]$. CL represents the state of being in one clause, and k is the number of words in tp_j , and kq is the number of decisive cue words in tp_j . Then, we define $f_g(tp_j) \in N$ as text pattern weighting scheme that if the $f_g(tp_j)$ is 1, TP containing tp_j is recognized as an idea and if the $f_g(tp_j)$ is 0.5, TP containing tp_j is a quasi-idea:

$$f_g(tp_j) = \begin{cases} 1 & | (\exists (j)(tp_j \in N) = \exists (i)(cp_i \in M)) \wedge (kq/k = 1) \\ .5 & | (\exists (j)(tp_j \in N) = \exists (i)(cp_i \in M)) \wedge (kq/k < 1) \\ 0 & | (\forall tp_j \in N) \neq (\forall cp_i \in M) \\ & (\forall i \in 1, \dots, \mu), (\forall j \in 1, \dots, \nu) \end{cases}$$

-- <Definition 1>

4.2 Cue Multimedia

A literary work, art object, and score has been made up with creator's inspiration and idea. A logical or mathematical formula like axiom, equation, etc, is also produced with researcher's trial of thinking newly and arranging logically. So, we can recognize the pictures, musics, equations, etc. as a idea related to human creation activities.

The figures and formulas are existed on web as a image file format like jpg, gif, etc. and the musics are as a sound file format like wav, mp3, etc. and the videos are avi, etc. file format. As it is described in the chapter 3, figures (such as images, pictures, photos, diagrams and icons), mathematical formulas (such as scores, equations, public interests and definitions), sounds and videos are generally able to be regarded as the outcome of ideas. But, certain forms of observation data should be excepted since they are simply observed data according to plans. Despite of that, they plays a role as preparation refernces in order to create ideas.

Thus, figures, formulas, sounds and videos emerging from documents and web-sites are extracted as a type of idea like as cue phrases of previous section. The extracting method is stated below and its applied regulation is the <Definition 2>.

- 1) Cue words revealing images, formulas, sounds and videos are distinguished.
- 2) Whether the images, formulas and tables are followed according to the cue words (contained in locations where the multimedia are indicated) is identified.

Definition 2. Let (a text) $\Theta = [\theta_1, \dots, \theta_n]$ be a list of general cue terms θ_i . Let $\Pi = [\pi_1, \dots, \pi_n]$ be a set of words π_j ($j \in [1, \dots, n]$) in a new text and multimedia stream Ω and let $\Sigma = [\sigma_1, \dots, \sigma_n]$ be a set of multimedia σ_i ($i \in [1, \dots, n]$) in Ω . Then, we define $P(\pi_j \in \Theta | \sigma_i \in \Sigma)$ as multimedia weighting scheme and, if the value of $P(\pi_j \in \Theta | \sigma_i \in \Sigma)$ is 1, Ω is recognized as ideas:

$$P(\pi \in \Theta | \sigma \in \Sigma) = \begin{cases} 1 | \prod_{i=1}^n \prod_{j=1}^n \frac{P(\sigma_i \in \Sigma | \pi_j \in \Theta) P(\pi_j \in \Theta)}{P(\sigma_i \in \Sigma)} \geq C & (\forall i, j \in [1, \dots, n]) \\ 0 | \prod_{i=1}^n \prod_{j=1}^n \frac{P(\sigma_i \in \Sigma | \pi_j \in \Theta) P(\pi_j \in \Theta)}{P(\sigma_i \in \Sigma)} < C \end{cases}$$

-- <Definition 2>

Where C is constant.

4.3 Contextual Signals

Contextual signal is the method that judges whether ideas are emerged according to searching contexts in documents. It assumes ideas when primary words

are emerged and depicting words in lists are appeared at the same time. And, if situation explanations about a certain topic and then signs about illustrations, assumptions, plans, reasons and solutions related to results, characteristics/circumstances and so on are emerged, it judges them ideas. The specific performing process is stated below and the <Definition 3-*> was applied.

- 1) We analyzed the characteristics of sentences from a new text stream and try to seek the contextual signature terms either a unemotional or a emotional.
- 2) When finding signature terms (referred <Table 3>) we calculate the cohesion degree of the paragraph included the term and then summate the weight of signature terms in the paragraph
- 3) As the value of adding the cohesion degree to the weight of signature terms is over the threshold, We recognize the present paragraph as an idea or a quasi-idea.

Definition 3: Let $\Theta = [\theta_1 \mu_1, \dots, \theta_n \mu_n]$ be a list of general signature terms θ_k and weights μ_k ($k \in [1, \dots, n]$). Let $\Psi = [\psi_1, \dots, \psi_n]$ be a list of signature terms ψ_k ($k \in [1, \dots, n]$) in S in P and the ψ_k be equal to the θ_k . Let $S = [s_1, \dots, s_n]$ be a set of sentences s_j

<Table 3> List of contextual signature terms

Unemotional	* suggest, suggestion, * express, expression, * announce, announcement, * present, presentation, * study, * research, * analyze, analysis, * create, creation, * design, * new, * axiom, * formular, * equation, * figure, * picture, * music
Emotional	* happy, * sad, * inspire, * love

($j \in [1, \dots, n]$) in a new paragraph stream P and let $P = [p_1, \dots, p_n]$ be a set of paragraphs p_i ($i \in [1, \dots, n]$) in a new text stream T . Let $CD = [\alpha_1, \dots, \alpha_n]$ be a set of cohesion degree α_i of p_i and let $SV = [sv_1, \dots, sv_n]$ be a set of signature value sv_i of p_i , which is a average value of sum of the signature terms weights in p_i . Let γ_k ($k \in [1, \dots, n]$) be a number of ψ_k or θ_k in p_i and $h = j$. We define $f_g(\alpha_i)$ as cohesion degree computing scheme where we use general cosine coefficient (Schutze, 1998) for calculating similarity degree, $Sim(s_h \in p_i, s_j \in p_i)$. And also define $f_g(sv_i)$ as signature value calculating scheme of paragraph p_i and $f_g(w_k, \psi_k)$ as signature term weighting scheme of ψ_k in p_i and $f_g(p_i)$ as paragraph weighting scheme.

$$f_g(\alpha_i) = \frac{1}{n(n-1)} \sum_h \sum_j Sim(s_h \in p_i, s_j \in p_i) \quad (\forall h, j \in [1, \dots, n])$$

-- <Definition 3-1>

$$f_g(w_k, \psi_k) = \log \frac{\gamma_k}{\mu_k} + \log \frac{\mu_k}{\gamma_k}$$

-- <Definition 3-2>

$$f_g(sv_i) = \frac{1}{n} \sum_m f_g(w_m, \psi_m) \quad (\forall m \in [1, \dots, n])$$

-- <Definition 3-3>

$$f_g(p_i) = \begin{cases} 1 & | f_g(\alpha_i) + f_g(sv_i) \geq 1 \\ 0.5 & | 0.5 \leq f_g(\alpha_i) + f_g(sv_i) < 1 \\ 0 & | f_g(\alpha_i) + f_g(sv_i) < 0.5 \end{cases}$$

-- <Definition 3-4>

When $f_g(p_i)$ is 1, we recognize the paragraph p_i as an idea and $f_g(p_i)$ is 0.5 the paragraph p_i as a quasi-idea.

4.4 Discourse Segmentation

This method allows that when expressions indicating ideas are emerged in conversations by each sentence searched one by one, the sentence will be accepted as ideas with the expressions increasingly emerged. Ideas were judged through the method in previous section 4.3. the other aspects different from section 4.3's are that dialogical sentences reacted as a judging target, and they were not only set up the <Fig. 3> as a outline to search, but the <Definition 4> were also applied.

Definition 4: Let $\delta_h \in \Delta$ ($h \in [1, \dots, n]$) be a list of significant word or phrase able to represented ideas. Let $t_i \in T$ ($i \in [1, \dots, n]$) be a significant word

*When 'cue words' intimating 'ideas' and cases of the words stated below are detected at the same time, they will be commonly configured as the 'candidate ideas'.

- 1) Words indicating solution, depiction, inference and so on
- 2) Words showing a sign of adaption towards other's persuasion.
- 3) Mentions of advertising, promoting and advertisement.
- 4) When reasons (experiencing a ghost hunt), irregular imaginations (it is cold in summer) and metaphorical (as if and like) are announced.
- 5) Words like "problem is that" or "in my opinion" in conversations.
- 6) When interrogative sentences are come up, weighted values are also added.

<Fig. 3> Significant words and phrases of dialogical sentences

or phrase in dialogue unit $d_i \in D$ ($i \in [1, \dots, n]$) from a new text stream T and w_i is a weight of δ_i . We define $f_g(ds_i)$ as a discourse segmentation weighting scheme of the d_i in present dialogue D .

$$f_g(w_i, t_i) = \log \frac{TF(t_i \in D)}{((w_i, \delta_i) \in \Delta)} + \log \frac{((w_i, \delta_i) \in \Delta)}{TF(t_i \in D)} \quad (t_i = \delta_i)$$

-- <Definition 4-1>

$$f_g(ds_i) = \frac{1}{n} \sum_i f_g((w_i, t_i) | (D \cap \Delta)) \quad \text{-- <Definition 4-2>}$$

$$f_g(p_i) = \begin{cases} 1 & | f_g(ds_i) \geq 0.5 \\ .5 & | 0.25 \leq f_g(ds_i) < 0.5 \\ 0 & | f_g(ds_i) < 0.25 \end{cases} \quad \text{-- <Definition 4-3>}$$

When $f_g(p_i)$ is 1, we recognize the paragraph p_i as an idea. and $f_g(p_i)$ is 0.5 the paragraph p_i as a quasi-idea.

5. Experimental Evaluation

The main purpose of this study has been to measure the extraction performance of four types of extracting methods stated in the section 4 according to below four problem domains:

- 1) The success rate of idea and quasi-idea extracting based on decisive cue phrase, contextual signal and discourse segmentation methods,
- 2) The success rate of idea and quasi-idea extracting when decisive cue phrase, contextual signal and discourse segmentation methods were applied into thought, plans, opinions, writing and textual formulas respectively,
- 3) The success rate of idea extracting of mathemat-

ical formula and multimedia methods,

- 4) The success rate of idea extracting when decisive cue phrase, contextual signal and discourse segmentation methods were mixed for the extraction.

When it came to the experiments, examples which were targets of the measurement had been divided into the form of text written by declarative sentences, the form of text written by conversational sentences, and the form of multimedia according to forms of the idea expressions. In the forms of declarative text, there were the total 120 number of examples which commonly included the text forms of “thought, plans, opinion and formulas” respectively; the 10 number of examples including ‘idea’, the 10 number of examples revealing quasi-idea and the 10 number of examples having ‘no-idea’. In the examples of the conversation form, there were writing examples which included idea, quasi-idea and no-idea in each the 10 number of example text. In terms of the form of multimedia, there were the total 60 number of the multimedia form on “figures, sounds and formulas”, and there were the 10 number of idea and the same number of no-idea in the examples.

The definition on classifying idea, quasi-idea and no-idea were constructed according to the established definition when examples had been chosen, and the other examples, which were not classified, were defined by five postgraduate students. Cue phrase and contextual signal among the four methods were measured with examples of the description forms targeting, and discourse segmentation and formula and multimedia targeted the forms of the conversation

and the multimedia respectively to experiment its performance. When the experiments were conducted, the judgement -that there were idea, quasi-idea and no-idea- had been defined by the other 10 postgraduates. At the stage, each experimenter suggested the average values of the results experimented (five times) by different search words respectively in the four extracting methods. And then the medians of precision and recall ratios statistically verified by median test with a single population was aligned into <Table 4, 5, 6, & 7> and the 'F Value' was calculated as well.

As the means to measure the extracting performance, the below F criterion was adapted, which is constructed by 'R (recall ratio)' and 'P (precision ratio)' mixed and it has been being generally used when recall ratio and precision ratio are not separately measured.

$$F = \frac{2PR}{P+R} \quad \text{--- F measure}$$

In this F value, the highest success rate was F value 1.0 when the precision ratio and the recall ratio were commonly 1.0. If the precision ratio and the recall ratio are in the situation of the inverse

proportion, it is theoretically normal that the recall ratio, the precision ratio and the F value are 0.5 in common. In this study, it was regarded that over 0.5 of the F values were assessed as positive results.

Firstly, the success rates of the idea and the quasi-idea extraction of decisive cue phrase, contextual signal and discourse segmentation methods had been measured through the examples, and then success rates of the extraction which are stated in the below <Table 4> were gained.

When the above three methods were individually adapted into thought, plans, opinions, writing and textual formulas, their success rates of the extraction on idea and quasi-idea is the same with the below <Table 5>.

(4) When decisive cue phrase and contextual signal methods had been mixed, its success rates were measured like <Table 6>. The measurement by individual example types is not noticed since the results were similar to the results of the <Table 4> and <Table 5>.

On the other hand, the success rates of mathematical formula and multimedia methods are stated in <Table 7>. In this table, the reason why the results of quasi-idea extraction were excepted is that there

<Table 4> The success rates of 'idea' and 'quasi-idea' extraction of reference analyzation by text forms

Method	Idea			Quasi-Idea			Number of Examples
	P	R	F	P	R	F	
Decisive cue phrase	.67	.71	.69	.25	.36	.30	120
Contextual signal	.35	.43	.39	.73	.70	.71	120
Discourse segmentation	.36	.34	.35	.62	.64	.63	30

<Table 5> The success rates of idea and quasi-idea extraction of reference analyzation by text forms according to types of the examples

Method	Form	Idea			Quasi-Idea			Number of Examples
		P	R	F	P	R	F	
Decisive cue phrase	Thoughts	.65	.73	.69	.29	.39	.34	120
	Plans	.70	.74	.72	.23	.38	.30	
	Opinions	.67	.71	.69	.28	.35	.31	
	Textual formulas	.68	.67	.67	.22	.32	.27	
Contextual signal	Thoughts	.37	.42	.39	.73	.68	.70	120
	Plans	.36	.46	.41	.77	.72	.74	
	Opinions	.35	.46	.40	.71	.67	.69	
	Textual formulas	.33	.39	.36	.72	.71	.71	
Discourse segmentation	Writings	.36	.34	.35	.62	.64	.63	30

<Table 6> The success rates of 'idea' and 'quasi-idea' extraction by the mixed method

Method	Idea			Quasi-Idea			Number of Examples
	P	R	F	P	R	F	
Mixed	.77	.82	.79	.82	.80	.81	120

<Table 7> The success rates of 'idea' extraction excepting the form of text data

Method	Form	Idea			Number of Examples
		P	R	F	
Multimedia	Figures	.41	.44	.42	20
	Sounds	.35	.36	.35	20
Mathematical Formula	Mathematical formulas	.55	.56	.55	20

were considerable professionals' views that distinguishing idea and quasi-idea was obviously difficult.

As a result of this study's experiments, the three methods except quasi-idea field of decisive cue phrase method and idea field of contextual signal and discourse segmentation methods can be regarded to be positive in the aspect of efficiency as F values

of the three methods, which dealt with texts, clearly exceeded 0.5 in the <Table 4>. Meanwhile, decisive cue phrase method showed better performance on extracting idea than quasi-idea in the <Table 4 and 5>. On the contrary, the rate of quasi-idea extraction was higher than the rate of idea in contextual signal method. This results indicate that the usage situation of above two methods should be different.

In the <Table 6>, the rate of the extraction was calculated applying decisive cue phrase and contextual signal method at the same time. As a result, applying mixed method was more efficient as it had shown higher efficiency than applying single method at about 10%. And why discourse segmentation method showed the lower percentage than decisive cue phrase and contextual signal methods at approximately 10% is that there had been relatively uncertain signature words which noticed idea written in sentences in the contents of discourse writings.

The extraction rate of the figures and sounds of multimedia methods were each 0.42, 0.35 in the <Table 7>. These relatively lower rates were considered to come from the different point of view between control group selected and experimental one distinguished ideas and quasi-ideas. And because there was observable tendency to much interindividual deviations in real-life situations. Meantime, recall and precision ratio is known to be in inverse proportion relationship to each other, but the property did not generally appear in this study. It is considered that the search language like as cue words and phrases, multimedia, and formulas are small in size and in synonym and homonym etc.

6. Conclusions

In order to extract idea and quasi-ideas from text stream inputted on blogs, on-line reviews, thesis, reports, and customer feedbacks on products etc., established methods were searched. Although the

established researches about idea mining were scarce, the techniques; decisive cue phrase, formula & multimedia, contextual signal and discourse segmentation, were commonly selected as the method of the ideas extraction among the techniques applied in opinion mining, automatic summarization and topic extraction methods since the methods were similar to the technique of idea mining. The ① technique was experimented on the assumption of automatic method entirely and the ②, ③, & ④ techniques were conducted on the assumption of semi-automatic method.

In this study, the types of the experimental examples had been configured ① thought, ② plans, ③ opinions, ④ writings, ⑤ drawings, ⑥ sounds and ⑦ formulas, and then text streams, which were belong to each type, were experimented with the text streams distinguished into ① idea, ② quasi-idea and ③ no-idea. At this moment, the number of each experimental sample was configured the 10 number. As a results of the performances of above four methods, there were positive effects at the examples of the text forms since over 50% of the success rates of the extraction were recorded in each method. And there was the increasing effect of the extraction at over 10% when the methods had been mixed and then applied. Thus, it is definite that applying mixed method is highly effective compared to applying single method. However, the reason of why the success rates were relatively low in discourse segmentation method was that there had been generally uncertain cue words used in discourse writing. Therefore, deeper researches in terms of the issue should be undertaken the next.

On the other hand, while the success rates of mathematical formulas reached 55% in examples of the un-textual forms, under 50% of the rates were shown in the forms of figures and sounds. So, since the efficiency in terms of illustrations, tables, diagrams and others should be measured, continuous researches on 'Index image', 'Index sound' and complementary signature term should be needed. In this study, decisive cue phrase method was considerably effective to search idea and contextual signal method was significantly effective to detect quasi-idea. Since quasi-idea is a definition to reach idea, the methodology to make idea should be reconsidered with quasi-idea using.

References

- Al-Halimi, R. K. (2003). Mining topic signals from text. Unpublished doctoral dissertation, University of Waterloo. Retrieved from: <http://uwspace.uwaterloo.ca/handle/10012/1165>.
- Barzilay, R., & Elhaadad, M. (1997). Using lexical chains for text summarization, In Proceedings of the Workshop on Intelligent Scalable Text Summarization at the ACL/EACL Conference, 2-9, Madrid, Spain.
- Bergstrom, T., & Karahalios, K. (2008). Conversation clusters: Human-computer dialog for topic extraction. Retrieved from: <http://social.cs.uiuc.edu/papers/pdfs/bergstrom-1361.pdf>.
- Brandow, R., Mite, K., & Rau, L. (1995). Automatic condensation of electronic publications by sentence selection. *Information Processing & Management*, 31(5), 675-685.
- Buitelaar, P., & Eigner, T. (2008). Topic extraction from scientific literature for competency management, Retrieved from: <http://citeseerx.ist.psu.edu/.../download?doi=10>.
- Businessdictionary.com (n.d.). Retrieved from: <http://www.businessdictionary.com/>
- Chung, Young Mee, & Kim, Yong Kwang (2008). A study on an effective event detection method for event-focused news summarization. *Journal of the Korean Society for Information Management*, 25(4), 227-243.
- Dave, K., Lawrence, S., & Pennock, D. M. (2004). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. Retrieved from: <http://www.kushaldave.com/p451-dave.pdf>.
- Definitions.net (n.d.). Retrieved from: <http://www.definitions.net/>
- Dey, L., & Haque, SK. M. (2008). Opinion mining fom noisy text data. Retrieved from: <http://dl.acm.org/citation.cfm?id=1390763>.
- Dong-A Daily News (2009). 2.19, 1.

- Edmundson, H. P. (1998). New methods in automatic extracting. In F. W. Lancaster, *Indexing and abstracting in theory and practice* (p. 269), London: Library Association Publishing.
- Hovy, E., & Lin, C. (1999). Automated text summarization in SUMMARIST. In *Proceedings of the Workshop on Gaps and Bridges in NL Planning and Generation*, 53-58. ECAI Conference. Budapest: Hungary.
- Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. *Proceedings of the Eighteenth Annual International ACM Conference on Research*, 68-73.
- Lee, Ji-Hye, & Chung, Young Mee (2009). An experimental study on opinion classification using supervised latent semantic indexing (LSI). *Journal of the Korean Society for Information Management*, 26(3), 451-462.
- Lee, Tae Young (2005). A Study on the construction of the automatic extracts and summaries: On the basis of scientific journal articles. *Journal of the Korean Society for Library and Information Science*, 39(3), 139-163.
- Liu, B. (2009). Opinion mining, Retrieved from:
<http://www.cs.uic.edu/~liub/FBS/opinion-mining.pdf>.
- Mani, I. (2001). *Automatic summarization*. Amsterdam: John Benjamins Publishing Company.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*, Cambridge, New York: Cambridge University Press.
- Marcu, D. (1999). Discourse trees are good indicators of importance in text. In I. Mani, & M.T. Maybury (Eds.). *Advanced in Automatic Text Summarization* (pp. 123-136). Cambridge, Massachusetts: The MIT Press.
- Meadow, C. T., Boyce, B. R., & Kraft, D. H. (2000). *Text information retrieval systems*. San Diego: Academic Press. 208-211.
- Myaeng, S. H., & Jang, D. H. (1999). Development and evaluation of a statistically-based document summarization system, In I. Mani, & M.T. Maybury (Eds.), *Advanced in Automatic Text Summarization* (pp. 61-70). Cambridge, Massachusetts: The MIT Press.
- Pang, B., and Lee, L. (2008). Opinion mining and sentiment analysis. Retrieved from:
<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.147.1344>
- Roth, B. (2007). Topic extraction and relation in instant messaging, Retrieved from:
<http://nlp.stanford.edu/courses/cs224n/2010/reports/rothben.pdf>
- Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1), 97-123.
- Teufel, S., & Moens, M. (1999). Argumentative classification of extracted sentences as a first step towards flexible abstracting. In I. Mani, & M.T. Maybury (Eds.), *Advanced in Automatic Text Summarization*(pp. 155-176). Cambridge, Massachusetts: the MIT Press.

The free dictionary (n.d.). Retrieved from: www.thefreedictionary.com/.

Thorleuchter, D. (2008). Finding new technological ideas and inventions with text mining and technique philosophy. Retrieved from:

<http://www.springerlink.com/content/j21800t0768x6644/>.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2009). Mining ideas from textual information. Retrieved from: http://www.feb.ugent.be/nl/Ondz/wp/Papers/wp_09_619.pdf.

Wang, X., Zhang, K., Jin, X., & Shen, D. (2008). Mining common topics from multiple asynchronous text streams. Retrieved from: <http://wsm2009.org/papers/p192-wang.pdf>.

Webster online dictionary (n.d.). Retrieved from: <http://www.websters-online-dictionary.org/>.

Wikipedia (n.d.) Retrieved from: <http://ko.wikipedia.org/wiki/>.

Yourdictionary.com (n.d.). Retrieved from: www.yourdictionary.com.