Web Contents Mining System for Real-Time Monitoring of Opinion Information

Ho-Bin Song[1], Moon-Taek Cho[1], Young-Choon Kim[2,*], and Suck-Joo Hong[3]

[1] Dept. of Electrical & Electronic Engineering, Daewon University, 599 Sinwol-dong, Jecheon, Chungbuk, 380-702, Korea
songhobin@paran.com, mtcho@mail.daewon.ac.kr
[2] Dept. of Car Engineering, Kongju National University, 275 Budae-dong, Seobuk-gu, Cheonam-si, Chungnam, 331-717, Korea
yckim59@kongju.ac.kr
[3] Dept. of Information and Telecommunication, Kyung Hee Cyber University, 1 Hoegi-Dong, dongdaemun-Gu, Seoul, 130-701, Korea
sjhong@khcu.ac.kr

Abstract. As the use of the Internet has recently increased, the demand for opinion information posted on the Internet has grown. However, such resources only exist on the website. People who want to search for information on the Internet find it inconvenient to visit each website.

This paper focuses on the opinion information extraction and analysis system through Web mining that is based on statistics collected from Web contents. That is, users' opinion information which is scattered across several websites can be automatically analyzed and extracted. The system provides the opinion information search service that enables users to search for real-time positive and negative opinions and check their statistics. Also, users can do real-time search and monitoring about other opinion information by putting keywords in the system.

Keywords: Motoring Search System, Opinion Information Automatic Extraction, Web Contents Mining, Opinion Information Monitoring.

1 Introductions

As the use of the internet gradually gets active lately, many people tend to express their opinions on the internet through the media such as Blog and Wiki[1]. And, in evaluating the value of the specific information, such demand for referring the opinion information other people put online is increasing. However, such opinions existing on the internet exist only at individual web site, and the user is to search all of such individual web sites manually in order to use such opinion informations. To settle this problem, technology for extracting the opinion of the user is actively being studied in home and foreign academics, and various technologies are being studied in

* Corresponding author.

T.-h. Kim et al. (Eds.): FGCN 2011, Part II, CCIS 266, pp. 43–56, 2011.

a field of information retrieval by great improvement from early 2000[1,3,5]. But the existing information retrieval technology is simply providing retrieval based on information having a keyword, and not being able to provide high-dimensional retrieval based on contents rated positively / negatively in document and sentence which each keyword appears. Recently an attempt to apply the technology for extracting the opinion of the user on information retrieval is in progress, but is in level of simply separating positive, negative document yet.

This thesis suggests web contents mining system for real-time monitoring of opinion information to settle this problem. Suggesting system provides opinion information information retrieval service able of retrieval and statistics in each positive/negative opinion by automatically extracting and analyzing opinion information of user from web contents scattered in many websites existing on internet. As a result, opinion retrieval users easily can use the system searching and monitoring opinion information of other users on the specific keyword readily at eye, and the function of automatically extracting and analyzing opinion information real-time in web contents is provided.

Construction of this thesis is as follows. In chapter 2, existing web mining technique, opinion extracting technique and theoletical background of multilingual linguistic dictionary are examined and the problem is examined for theoletical inquiry of this thesis. In chapter 3, plan and design method of web contents mining system able to collect and analyze opinion information on the internet is suggested. Finally in chapter 5, conclusion Is formed.

2 Relevant Study

2.1 Web Mining Technique

Web Mining is aimed at all data originated on the web or existing on the internet, and indicates the process of extracting and analyzing useful information by applying data mining technique based on such data. That is the application which data mining technique is applied to the web, massive data assembly [2,4,5]. Such web mining uses data mining technique to find and extract information automatically from web document and service. So, it can be defined as the process of finding useful information and knowledge not known previously from web data.

Field of study of web mining is holding in common much parts studied in the field of Information Retrieval or Information Extraction [3,4,5].

2.2 Opinion Extracting Technique

This is the study of opinion classification in which unit of document and sentence is classified more in detail into unit of phrase and word. Study of classifying opinion at the unit of phrase and word is studied by the method based on rule at the beginning, and the machine learning method of studying information around phrase and word and deciding polarity of phrase and word is studied afterward [7,8].

2.2.1 Rule-Based Model

Study of Nasukawa and Zhongchao Fei is done as the study of extracting opinion in unit of phrase based on rule-based method [9,10]. Part of speech information and polarity information of each word are put together with that word and tag is attached [11,12]. Polarity information is determined to three, good, bad, neutralm, and word part of speech being object is determined to adjective, noun, adverb and verb. Tag is attached on positive verb and negative verb. At this point, tag is attached on one pertinent word or also on phrase forming one expression with that word and showing specific polarity.

2.2.2 Machine Learning Based Model

It was mainly focused on the method of manually constructing opinion expressing resource at previous rule-based method. Machine learning is carried out using Corpus tagged to positive or negative expression part in sentence on the machine learning method.

After automatic construction of tagged Corpus, the machine learning for opinion classification of word / phrase unit using this Corpus is done. HMM (Hidden Markov Model) is used as machine learning for opinion classification [16].

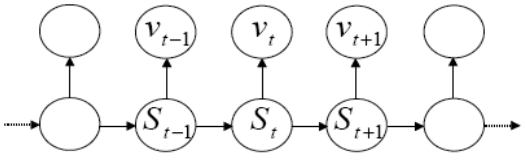


Fig. 1. Probabilistic Parameters at HMM

2.2.3 Pros and Cons of Prior Method

Prior rule-based method has merit of showing high accuracy on corresponding pattern. So it can be the good method to approach if many opinion patterns could be constructed delicately. But, as the problem the rule-based method was constantly being pointed out, in case of rule-based method, there is weak point of which reproducibility drops sharply in case pattern already constructed come out in transformed form, and there is limit of not being able to reflect surrounding context information to studying. Constructing and maintaining this opinion pattern on all other domains and linguistic range is the hard work necessary of much manpower and time.

There is difficult which opinion resource is to be constructed manually in both rule-based method and machine learning based method for opinion classification of word / phrase. For the opinion classifying of word / phrase unit is possible if such opinion resource is constructed, one of the most important problems in opinion classifying of word / phrase unit could be seen as construction of Corpus which opinion pattern construction or opinion expression is tagged.

2.3 Multilingual Linguistic Dictionary

2.3.1 Foreign Words Automatic Transcribing Model

Study on automatically constructing translation knowledge used in application of natural language such as machine translation and information retrieval of cross lanugage actively has been progressed [6,7].

Phonetic translation means of generally transcribing English word in language of non-English-area based on pronunciation. Many of the words phonetic translated are not registered in dictionary for many of them are coined words showing new ideas. Therefore automatically obtaining translation knowledge of phonetic translation is very important to build effective translation knowledge. We have automatic phonetic translation and phonetic translation interlinear pair extraction, etc. as the study of obtaining phonetic translation interlinear words on given English words. Automatic phonetic translation is the technique of phonetic translating given English word into a word of non-English-area [7].

Phonetic translation interlinear pair extraction, the study of automatically extracting English and phonetic translated word corresponding to English from the form bilingual corpora to widen applicative range of translation dictionary, and the translation knowledge is limited to phonetic translation interlinear pair. Automatic phonetic translation and phonetic translation interlinear pair extraction are actively being progressed as the method of handling phonetic translated word, but the study of integratively using these two methods is not thoroughgoing enough.

2.3.2 Statistics Based Phonetic Translation Model

Generally Roman notation of Chinese characters or Pinyin is used in comparing with English in Chinese phonetic translation interlinear pair extraction [15]. With assumption of E as English, C as Chinese, TU (Translation Unit) as phonetic translation unit in statistics based phonetic translation model, conditional probability P(C|E) is substituted with P(Chinese|English) and can be converted to the problem of seeking P(C|E) probability. And, Unigram, Bigram, Trigram for English, and Pinyin's first syllable, last syllable or the whole Pinyin for Chinese are used as TU.

Method of automatically presuming parameter by applying EM(Expectation Maximization) algorithm [13] without pronouncing dictionary is used, and match type information is added to phonetic translation model. With adding match type (M) to P(C|E) formula, we have formula 1.

$$P(C|E) \fallingdotseq \max P(C|M,E)P(M|E) \fallingdotseq \max P(C|M,E)P(M) \qquad \text{(formula 1)}$$

3 Web Contents Mining System for Real-Time Monitoring of Opinion Information

3.1 Outline of System

Web contents mining system for real-time monitoring of opinion information is the system to automatically extract and analyze opinion information in web contents, and

that platform is as figure 2. Proposed system is the system that provides opinion information searching service able to check search and statistics in each positive/negative opinion by automatically extracting and analyzing user opinion information from web documents scattered over many websites existing on the internet. Positive opinion and negative opinion are automatically extracted.

Proposed system of figure 2 is formed by including data collection processing, opinion/non-opinion automatic construction, Opinion information resources, indexing transaction, opinion indexing information resources, opinion expression machine learning, multilingual linguistic dictionary automatic registration, multilingual opinion information resources, opinion search transaction and user terminal, etc. are included in forming.

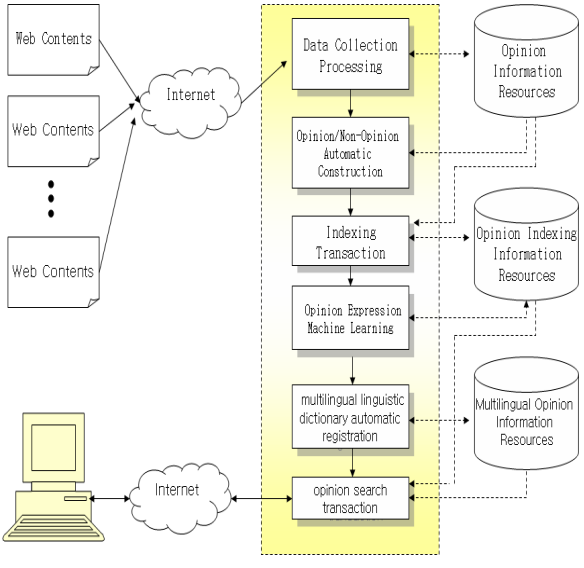


Fig. 2. Opinion information automatic extraction web mining system platform

3.2 Data Collection Processing

Data collection processing performs function of collecting various web contents existing on the internet. So, data collection processing downloads HTML(Hyper Text Markup Language) information of each Web Site existing on the internet in real time. Moreover, data collection processing can extract information data of at least one among necessary information such as text, image or video, etc. from web contents downloaded as above and store in separate data storing module.

Data collection processing can sort and collect web contents including opinion information data (that is general sentence/document data and information data with positive/negative valuation on it) as table 1. Object data collected through data collection processing as shown in table 1 is the opinion information data that is the general sentence/document data and the information data with positive/negative

valuation on it. At this time, the above positive/negative valuation can be expressed in point within fixed range or variously valued using the asterisk(★) or other marks. Positive/negative valuation so expressed in various methods are all recalculated in same point range and used in this thesis.

Table 1. Opinion Information Data

Expression	Point	Opinion Contents
★★★★★	10	Interesting. statement
★★★★★	10	Story of 'smart' people. statement
★★★★☆	8	Wise people mending ordinary life! statement
★★★★★	9	Fascinated by uncle ... statement
★★★★☆	8	Story of ordinary people. statement
★★★★★	10	heart warming love story with spelendid acting and interesting story. statement
★★★★★	10	Really touching story. statement
★★★★★	10	Heart warming movie. Interesting also, statement
★★★☆☆	6	Warm and funny ... Could be longer ... statement
★★☆☆☆	5	palpable story after all. statement

To explain this in the concrete, if the point range of collected data is c ~ d when the point range used in working example of this thesis is a ~ b, the pertinent collection point x is changed as formula 2.

$$PolarityScore(x) = (a-1) + \frac{x-c+1}{d-c+1} \times (b-a+1) \quad \text{(formula 2)}$$

For instance, in case this thesis uses the point between 1 ~ 10 point (positive as closer to 10 point), and the collected data uses the point between 1 ~ 5 point, if the collected data is 2 point, it is calculated as formula 3.

$$PolarityScore(2) = (1-1) + \frac{2-1+1}{5-1+1} \times (10-1+1) = 4 \quad \text{(formula 3)}$$

Data collected by data collection processing is stored in opinion information resources data structure like table 2 to express in the set of opinion point {(data, point), (data, point), ... (data, point)} used in this thesis like chart 1. Table 2 is showing field name of opinion information resources data structure, data type on pertinent field and explanation on field.

Table 2. Opinion Information Resources Data Structure

id	user_id	date	topic	sentence	polarity	.....
bigserial (PK)	char varying(200)	bigint	char varying(200)	char varying(1000)	char varying(10)	.....

This thesis has automatically building study corpus which opinion information "word/phrase" is tagged as an object. Thus opinion information "word/phrase" is automatically classified through machine learning method using corpus automatically built. At this time, data collection processing collects data which positive/negative opinion information in blocks of sentence easily sought in the internet is expressed using opinion information data structure of formula 2 and table 2 to automatically built corpus that opinion information "word/phrase" is tagged.

3.3 Opinion/Non-opinion Automatic Construction

Method of simply using the number of times appealing in positive/negative document based on rule like formula 2 is inaccurate on data in form of point like 1~10 point. And in case of using the absolute number of times appealed when the number of positive document and negative document are different, there is the problem of having the point leaned to the pertinent data set collection of bigger size.

Method of automatically constructing opinion/non-opinion information resources in this clause has feature of automatically seeking through interpolation the positive/negative probability and the probability of appearing in opinion sentence of candidate word to be used after generating possible opinion information expressions by the whole dictionary based N-Gram analyzers. Opinion intensity of various point collection like 1~10 point is reflected in process of seeking positive/negative probability and probability of appealing in opinion sentence, and normalization is also proposed to settle the problem of point leaning of data size itself of specific point collection has grown.

3.3.1 Word Point Calculating Method of Proposed Method

Data which opinion is indicated in sentences is used in this clause to automatically construct opinion information word resource. After that, the point on N-Gram of each morpheme is sought after dividing sentence into morphemes.

$Freq(W_i, S_i)$ shows the number of times the word W_i appeals in the point collection Si. So supposing the point of 10 appeared 9 times $Freq(movie, s_{10}) = 9$ and the point of 1 1 time $Freq(movie, s_1) = 1$ for the word "movie", the extreme point using the frequency on "movie" in positive sentence and negative sentence is as formula 4.

$$Score(Movie) = \log\left[\frac{Freq(Movie, s_{10})+1}{Freq(movie, s_1)+1}\right] = \log\left[\frac{9+1}{1+1}\right] = 1.60 \quad \text{(formula 4)}$$

Formula 4 is the example of which problem arises when former method of clause 2.2 is directly applied. It is used under the condition in which the sizes of positive/negative data were same at the former study. Problem arises if the size of each data is calculated as formula 4. In case the word “movie” appeared 9 times at 10 point collection and 1 time at 1 point collection, very positive point of 1.6 is won by simply calculating with the above numerical formula. But as we see the example above, we can see that each word appears more at 10 point collection for 10 point collection itself is big. So the problem arises in case size of each point collection varies in seeking polarity point of the word.

As the example above, the word "movie" is the word that appears a lot in all point collections in common at the movie review. At this time, the problem of which the point is excessively biased to big point in case the absolute value simply is used in the condition which the absolute size of 10 points itself is big. Therefore, normalization without simply taking average of point is needed in case size of each point collection varies. Formula 5 is the numerical formula for converting absolute frequency to relative probability.

$$Freq(w_j, s_i) \rightarrow \frac{Freq(w_j, s_i)}{\sum_{w_k \in W} Freq(w_k, s_i)} = P(w_j | s_i) \quad \text{(formula 5)}$$

On formula 5, $Freq(W_i, S_i)$ indicates the number of times the word W_j appears at the point collection Si, $\sum_{w_k \in W} Freq(W_k, S_i)$ means the value that added the number of times the word W_k appears at all point collections, after all the number of times W_k appears in the whole data. So $P(w_i | s_i)$ is the value which absolute frequency is converted to relative probability. Normalizing absolute frequency to relative value having relative probability $P(w_i | s_i)$ based on this comes to formula 6.

$$PolarityScore(w_j) = \frac{\sum_{s_i \subset S} [i \times P(w_j | s_i)]}{\sum_{s_i \subset S} P(w_j | s_i)} \quad \text{(formula 6)}$$

Normalized opinion point not leaned to specific point collection is obtained by changing the part which had the average of absolute value formerly to the average of normalized values according to formula 6.

3.3.2 Calculating Method Using Subjective Point

Subjectivity of word as well as polarity point should be calculated in constructing word resources. At this time, the element having an influence on subjectivity as calculating polarity point is to be the generative probability of part of speech information of that word rather than the generative probability of that word itself. It is because of being weak to various proper noun, naturalized word and other words not in study data in case of generative probability of word itself, and that part os speech information can be usefully used for the specific part of speech combination exists. Subjectivity of word can equally be calculated using the method of calculating point of word previously proposed, but the subjectivity point of that word can be sought by calculating in seeing calculating subject data as the part of speech data of opinion data and the part of speech data of non-opinion data as positive(10 point) and negative(1 point) data each.

Formula 7 is the method of calculating subjective point in which pos_j indicates part os speech information of word w_j, and s_i indicates each point collection. s_1 is seen as the data collection not including opinion and s_{10} is seen as the data collection including opinion in calculating subjective point.

$$Score(pos_j) = \frac{\sum_{s_i \subset S} [i \times P(pos_j | s_i)]}{\sum_{s_i \subset S} P(pos_j | s_i)} \quad \text{(formula 7)}$$

Formula 8 is the proposal method automatically constructing opinion information word of this thesis which seeks the Opinion Score of word using two points, polarity point and subjective point, and using interpolation method. This Opinion Score is the point showing how much opinion word that word is, thus the word under specific point is not used as the opinion word.

$$\begin{aligned} &OpinionScore(w_j) \\ &= \left| PoarityScore(w_j) - \frac{1}{2} \times \max(S) \right| \times \alpha \\ &+ \left[SubjectiveScore(pos_j) - \frac{1}{2} \times \max(S) \right] \times (1-\alpha) \end{aligned} \quad \text{(formula 8)}$$

max(S) of the above numerical formula means the maxium point collection from the point collection. Reason of having absolute value on $-\frac{1}{2} \times \max(S)$ part and Polarity Score part is to have negative words heighten opinion point also at Polarity Score.

3.4 Opinion Expression Machine Study

After opinion/non-opinion tagging corpus is automatically constructed, the machine learning for opinion classification in word/phrase is done. HMM described in relevant study of chapter 2 is used as machine learning for opinion classification. But there are evaluation problem, decoding problem and estimation problem to be settled for HMM to be actually applied [16].

Evaluation problem to be settled first is the problem of how to seek the probability $P(O \mid \lambda)$ of data O observed in the model when sequence $O=O_1O_2...O_r$ and model $\lambda = (A, B, \pi)$ of the observed symbol is given. Decoding problem to be settled second is the problem of what the optimum state transition sequence $Q = q_1 q_2 \cdots q_t$ is when sequence

$O=O_1O_2...O_r$ and model $\lambda = (A, B, \pi)$ of observed symbol is. Estimation problem to be settled third is the problem of deciding model parameter $\lambda = (A, B, \pi)$ showing the biggest $\pi = \{\pi_i\}$.

Three problems above can be settled by Forward Algorithm, Viterbi Algorithm and Baum-Welch Algorithm each. In this thesis, state transition probability, observation probability and initial state probability which is the model parameters are obtained from the tagging corpus collected from opinion/non-opinion automatic construction module.

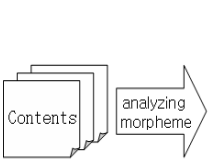
3.5 Indexing Transaction

Indexing transaction performs the function of indexing for the opinion informations of pertinent web contents to be stored in opinion indexing information resource in linguistic qualities of opinion sentence sorted from opinion/non-opinion automatic construction. Opinion indexing information resource here performs the function of summary information of relevant opinion sentence of linguistic qualities of each opinion sentence indexed through indexing transaction and basic and opinion informations of relevant web contents to be stored as database.

Table 3. Opinion Indexing Information Resource Data Structure

id	comm entct	date	snippet	data	pola rity	topic	url	...
big serial (PK)	char varying (50)	bigint	char varying (200)	char varying (2000)	char varying (10)	char varying (10)	char varying (200)	...

Opinion indexing information resource which opinion informations of pertinent web contents are to store in linguistic qualities classified from opinion/non-opinion automatic construction is stored in data structure like table 3. Table 3 shows the explanation on field name of opinion indexing information resource data structure, data type on pertinent field and field.



Large Classification Table

Index key	Content Number	Freque -ncy	Title Yes/No
Movie	53	12	1
IT	85	6	1
Movie	97	48	0
Movie	11	4	0
Environ -ment	99	2	0

Fig. 3. Large Classification Indexing

Indexing process is constituted of the large classification indexing process to improve search speed and the detailed classification indexing to use in actual information search process by indexing guide word and contents information of each document. Large classification indexing shows the document including technical terms. Figure 3 below shows composition of large classification index.

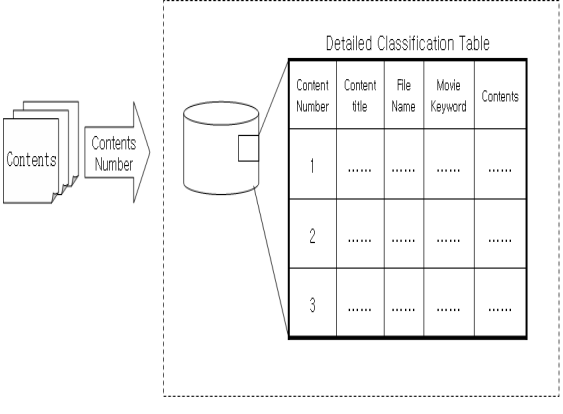


Fig. 4. Detailed Classification Indexing

Detailed classification indexing is the process of forming document table to search actual documents including the keyword user presented. Information such as subject information, File name (storing channel), document contents and major related keyword included in document contents are stored in document table. Movie related keywords are extracted and stored through inquiring movie review sentence in analyzing morpheme here. Figure 4 below shows composition of detailed classification indexing.

3.6 Multilingual Linguistics Dictionary Automatic Registration

In this clause, Effect of drastically reducing human power compared to existing passive linguistics dictionary constructing method is obtained by proposing the method of automatically constructing multilingual linguistics dictionary at double language on the internet using statistics based phonetic translation model.

This clause suggests the phonetic translation interlinear pair extraction method at mass comparative corpus using phonetic translation frequency and phonetic similarity based on dynamic window and tokenizer technique applied to parallel corpus.

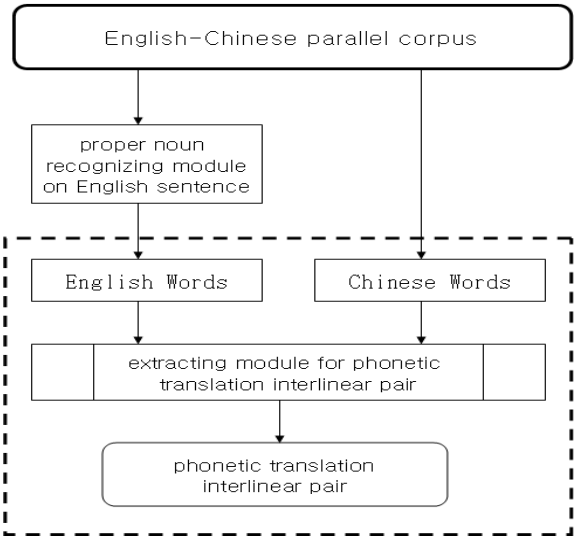


Fig. 5. Phonetic Translation Interlinear Pair Extraction Process in English-Chinese Corpus

English-Chinese phonetic translation automatic extraction model proposed in this clause first extracted proper noun applying proper noun recognizing module on English sentence of English-Chinese parallel corpus, chose only English words to be phonetic translated among them and extracted phonetic translation word from corresponding Chinese sentence. Figure 5 shows the process of extracting phonetic translation interlinear pair from English-Chinese parallel corpus, and the parallel corpus data structure is as table 4.

Table 4. Multilingual Opinion Information Resource Data Structure

id	from language	to language	from_text	to_text	...
bigserial (PK)	char varying(50)	char varying(55)	char varying(2000)	char varying(2000)	...

As observed in related study of chapter 2, error frequently occurs in extracting phonetic translation interlinear pair applying statistics based phonetic translation model if Chinese string similar in pronunciation with English word given in one sentence exist a lot. This thesis proposes dynamic window technique and tokenizer technique, non phonetic translation technique using entropy, and phonetic translation extraction technique using similarity and frequency of voice to settle error.

3.7 Opinion Information Search Transaction

Opinion information search transaction provided of specific opinion information or type information of user transmitted through web server and linked with indexing transaction or opinion indexing information storing resource, carries out function of searching indexing informations related with specific opinion search keyword or type information, and forwarding to web server to be transmitted to pertinent user.

4 Conclusion

Opinions existing on the internet exist only in individual web sites, so the user has to search such individual web sites one by one manually in case of using such opinion informations. Web contents mining system for real-time monitoring of opinion information is proposed in this thesis to settle such problems. Proposed system provides opinion searching service that can check retrieval and statistics in positive/negative opinions by automatically extracting and analyzing user opinion informations from web contents scattered in several web sites existing on the internet.

Expected effect of the proposing system is that the users are able of easy and at a glance searching and monitoring of opinion information of other uses on the specific keyword and the time spent for searching opinion of other uses can be greatly shortened by automatically extracting and analyzing user opinion informations scattered in several web sites existing on the internet and providing opinion searching service to be able to check searching and statistics in positive/negative opinions.

As the tasks to be solved, the web contents opinion searching system for the complete monitoring search engine is to be made by adding multilingual (Korean, Chinese, Japanese, English) search and machine translation function to solve language barrier on the internet and to be able of monitoring foreign informations in native language, and the reliability on opinion monitoring is to be verified.

References

[1] Joo, H., Park, Y.: Design of Web Contents Mining System for Monitoring Search Engine. In: KICS, vol. 34(2), pp. 53–60 (February 2009)
[2] Jang, N., Hong, S., Jang, J.: Data Mining. DaeChung, 32–56 (2007)
[3] Anand, S., Bell, D., Hughes, J.: The Role of Domain Knowledge in Data Mining. In: CIKM (1995)

[4] Anand, S., Hughes, J.: Hybrid Data Mining Systems: The Next Generation. In: PAKDD 1998, pp. 13–24 (1998)
[5] Adriaans, P., Zantinge, D.: Data Mining. Addison Wesley Longman, England (1996)
[6] Berry, J., Linoff, G.: Data Mining Techniques: For Marketing, Sales, and Customer Support. John Wiley & Sons (1997)
[7] Kosala, R., Blockeel, H.: Web Mining Research: A Survey. In: ACM SIGKDD (July 2000)
[8] Lee, C.H., Yang, H.C.: A Web Text Mining Approach Base on Self-Organizing Map. In: Proceedings of the 2nd International Workshop on Web Information and Data Management, WIDM 1999, Kansas City, MO, USA, pp. 59–62 (1999)
[9] Mulvenna, M., Anand, S., Büchner, A.: Personalization on the Net using Web Mining. Communications of the ACM 43(8) (August 2000)
[10] Dagan, I., Church, K.W., Gale, W.A.: Robust bilingual word alignment for machine aided translation. In: Proceedings of the Workshop on Very Large Corpora, pp. 1–8 (1993)
[11] Lee, J.S., Choi, K.S.: Enflish to Korean Statistical transliteration for information retrieval. Journal og Computer Processing of Oriental Languages 12(1), 17–37 (1998)
[12] Kang, B.J., Choi, K.-S.: Automatic Transliteration and Back-transliteration by Decision Tree Learning. In: Proceedings of LREC (2000)
[13] Goto, I., Kato, N., Uratani, N., Ehara, T.: Transliteration Considering Context Information Based on the Maximum Entropy Method. In: Proceedings of MT-Summit IX (2003)
[14] Yan, Q., Grefenstette, G., Evans, D.A.: Automatic transliteration for Japanese-to-English text retrieval. In: Proceedings of ACM SIGIR 2003, pp. 353–360 (2003)
[15] Paola, V., Paola, V., Khudanpur: Transliteration of Proper Names in Cross-Lingual Information Retrieval. In: ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition (2003)
[16] Dorre, J., Gerstl, P., Seiffert, R.: Text Minin g:Finding Nuggets in Mountains of Textual Data. In: Dorre, J., Gerstl, P., Seiffert, R. (eds.) Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (1999)
[17] Yan, L.H.: Text Mining-Knowledge Discovery from Text. In: Trend in Knowledge Discovery from Databases (June 29, 1999)
[18] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery. In: Advances In Knowledge Discovery and Data Mining, pp. 1–34. AAAI Press/MIT Press, CA (1996)
[19] Sproat, R., Tao, T., Zhai, C.: Named Entity Tranliteration with Comparable Corpora. In: Proceddings of the 21st International Conference on Computational Linguistics (2006)
[20] Oh, J.-H., Bae, S.-M., Choi, K.-S.: An Algorithm for extracting Engilish-Korean Transliterationpairs using Automatic E-K Transliteration. In: Proceedings of Korean Information Science Society (2004)
[21] Lee, C.J., Chang, J.S., Jang, J.S.: Extraction of transliteration pairs form parallel corpora using a statistical transliteration model. Information Science 176, 67–90 (2006)
[22] Lee, C.-J., Chang, J.S., Roger Jang, J.-S.: Alignment of bilingual named entities in parallel corpora using statistical models and multiple knowledge sources. ACM Trans. Asian Lang. Inf. Process 5(2), 121–145 (2006)
[23] Satish, L., Gururaj, B.: Use of hidden Markov models for partial discharge pattern classification. IEEE Transactions on Dielectrics and Electrical Insulation (April 2003)