http://dx.doi.org/10.5392/IJoC.2012.8.4.021

Heterogeneous Web Information Integration System based on Entity Identification

Hyung-Wook Shin
Development Inzisoft Co.LTD Seoul 135-080, South Korea

Hyung-Jeong Yang, Soo-Hyung Kim, Guee-Sang Lee
Department of Computer Science
Chonnam National University, Gwangju, 500-757, South Korea

Kyoung-Yun Kim
Department of Industrial and Manufacturing Engineering
Wayne State University, Detroit, MI 48202, USA

Sun-Hee Kim
Department of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213, USA

Do Luu Ngoc
Department of Computer Science
Chonnam National University, Gwangju, 500-757, South Korea

ABSTRACT

It is not easy for users to effectively have information that is semantically related but scattered on the Web. To obtain qualitatively improved information in web pages, it is necessary to integrate information that is heterogeneous but semantically related. In this study, we propose a method that provides XML-based metadata to users through integration of multiple heterogeneous Web pages. The metadata generated from the proposed system is obtained by integrating different heterogeneous information into a single page, using entity identification based on ontology. A wheelchair information integration system for disabled people is implemented to verify the efficiency of the proposed method. The implemented system provides an integrated web page from multiple web pages as a type of XML document.

Key words*: Semantic Web, Ontology, OWL, Integrating Information, Meta Data.*

1. INTRODUCTION

Although the Internet provides a convenient way to search information, an explosive increase of information engenders people to spend great amount of time and effort to obtain the information they need. Therefore, it is necessary to build a system for obtaining qualitatively useful information for users from web pages. The Semantic Web is considered as a solution for the next generation web [1]. The Semantic Web provides useful information to users by judging semantics between resources and the information of resources such as Web documents, files, and services. This is a framework where ontology is used to provide knowledge that machines understand. However, it is still necessary to establish a method that provides integrated information for similar information across different places. For example, suppose we want to retrieve information for a product and search engines will show endless of web links where similar information is included yet with different terms or units.

Metadata plays an important role that is able to describe different heterogeneous information through an integrated format using entity identification. Metadata is structured data

* *Corresponding author, Email: hjyang@chonnam.ac.kr*
Manuscript received Sep. 17, 2012; revised Nov 30, 2012; accepted Dec 10, 2012

about a document itself. It includes network addresses and various access points which exist independent from resources [2]. More specifically, metadata is descriptive information about a resource (physical or an electromagnetic) or an object. For instance, a generated record by index or a list by title, author, subject, classify symbol, etc., in a library, can be metadata in these means. The combination of the Semantic Web and metadata becomes a noble approach for configuring systematic documents for developers. In addition, a description method for the documents that uses systematic document structures and specific tags would help users find required information quickly and accurately. Thus, studies on the configuration of metadata using the Semantic Web have been largely conducted for achieving more advanced searching [3].

Web information retrieval expands searching power from a limited retrieval of targeting possession data to targeting electron resources. However, the information presented on the web is composed in various semantic languages in different categories, and furthermore, different words are used despite the fact that they belong to the same category. Also, due to multiple definitions of certain words, search engines present difficulties in its information search. Therefore, a method that can identify homonyms and synonyms as well as provide an integrated format to users through heterogeneous data integration is required [4]. New information retrieval methods based on ontology, which semantically connects knowledge concepts and effectively manages web resources in various web environments, are required.

In this study, web pages are classified using pre-learning and specific tags as a primary step. They are then dispatched into categories by identifying their homonyms and synonyms using an ontology data dictionary. Next, a method that provides the integrated XML metadata, established by extracting name entities and values from scattered heterogeneous data, is proposed. In addition, a wheelchair information integration system for disabled people is implemented to verify the efficiency of the proposed method.

This study consists of five sections. In Section 2, background studies on the configuration of metadata and information integration methods are examined. In Section 3, a system that configures metadata for implementing information integration is proposed. In Section 4, the proposed system is implemented and the efficiency of the system is measured. In Section 5, the conclusion and future studies are presented.

2. RELATED WORKS

Due to the expectation about technology developments in the information retrieval field, the Semantic Web has been examined for providing various retrievals and accuracy to users since computers are supposed to understand the human language and suggest an answer about a query in the semantic environment. Metadata, ontology, and range selection of search word are considered in the Semantic Web [5][6]. Moreover, metadata has been studied using various methods to consider the significant phase in supporting the semantic retrieval.

By adding a category called the Semantic Web, Nate [7], a portal site, established metadata from information associated to search keywords input by users. Conveniences along with interesting information are provided to users by combining the user search results with Social Network Services (SNSs). Although this method represents an advantage in addressing various results for a specific keyword, it may require a great deal of time than conventional searches because this method presents an overabundance of search results. Since this semantic search method provides documents that are considered to be related to the given keyword as links, similar to the conventional search engines, users must accept the inconvenience of having to verify the content of the documents one by one via tracing the links [4].

In cases of the other portal companies, such as Microsoft Bing, Google, and Yahoo provide the method consisting of metadata grafting the Semantic Web. The methods for consisting the metadata are Data Grid [8][9], Data Integration [10][11], and Unified search based on natural-language analysis. Yahoo SearchMonkey [12] uses Relational Database(RDB)[13] that applies the Semantic Annotation. The Semantic Annotation is a method that uses a type of comment which represents the category of a subject document by applying a specific tag to the document. This method usually has been used to convert unstructured documents to structured documents. However, it is not possible to apply this method as systematically configured RDB is not provided. Also, it shows a limitation in providing information because the user participation is not considered, although such RDB is established.

A metadata based data grid system [8] for integrating military information provides integrated military information based on heterogeneous databases. The system provides uniformed services in order to represent, store, and access different types of metadata. The scattered data are integrated as a single piece of information through a metadata manager, and the integrated data will be expressed on a user interface through a query processor. However, this method provides information according to the configuration of databases regardless of users' intentions. In addition, it does not provide a process for homonyms and synonyms.

In order to obtain the qualitatively improved results from web pages, scattered data which has an interrelated signification should be integrated though heterogeneity. The existing methods on intelligence fusion portray major problems in respect to the limitation of accessibility and expendability. These problems come from the database that is not user accessible. Namely, information fusion based on database is accessible only by the administrator. Therefore, the ordinary users are prohibited to access and operate information in database to have integrated information.

To solve this problem, [8] proposed a method that establishes the metadata using ontology, and automatically analyzes the semantics and relations of data. This method extracts information in the preprocessing step and is trained using Hidden Markov Model (HMM). Finally, ontology is used to determine semantics. However, this approach is focused on determination of the ambiguity of semantics than configuration of metadata. Corcho [9] proposed the Semantic Open Grid System Architecture (S-OGSA) to construct metadata. This method configures metadata that is robust to homonyms and synonyms based on RDF [14][15] and Web Ontology

Language (OWL) [16]. It constructs metadata through different processes according to natural language-, tag-, and RDF-based environments. However, this method does not present practical configuration processes or the results of the experiments.

In this study, we propose a system that solves the problems as mentioned above. We establish metadata based on ontology and provide integrated data to users as a form of XML document. In addition, a wheelchair information integration system for disabled people is implemented for verifying the efficiency of the proposed method.

3. INFORMATION INTEGRATION SYSTEM

The proposed system for integrating heterogeneous data in web pages consists of four phases: preprocessing, information extraction, reliability evaluation and information integration. Firstly, probability learning is performed using Bayesian Theorem in order to compute features which are related to a given domain in the preprocessing step. Data dictionary with ontologies is constructed to obtain related terms, homophones and synonyms. This data dictionary is used to construct metadata by means of name entity.

Secondly, the process of information extraction removes unnecessary information such as redundant and duplicated words according to the user input keyword from the retrieved web page. This step can efficiently reduce memory usage by deleting unnecessary data as well as increase the speed of the system. Thirdly, reliability decision analyzes that each data belongs to any category according to the pre-defined rules. Finally, information integration is performed using metadata to improve retrieval results. In this study, an independent system from databases is established since all data are to be configured as independent files of XML and OWL formats. In addition, the inconvenience of the conventional search method, needing to re-check the search results, is improved by immediately providing search results in an integrated form instead of presenting the results as links.

3.1 Preprocessing

Preprocessing is divided into pre-learning and data dictionary establishing steps. The pre-learning process is usually used in classifying patterns. Bayesian Learning Algorithm establishes a model for both useful and non-useful data, through repetitive processing using the training data. The model can reduce the complexity in its calculation through screening unnecessary data such as noise so that this model boosts system performance. The learning module is carried out by utilizing a prepared training set based on the Bayesian Theorem [17]. In learning module, HTML tags included in the training set are removed followed by a deletion process for stop words which represent words that do not affect the search results even though they appear frequently in documents.

Articles and prepositions in English and postpositions in Korean are considered as stop words. Since terms used in documents are represented in variations on its spelling according to certain variables, such as past tense, future tense, singular, and plural, the calculation complexity is significantly increased if a stemming process is not used. The JLex (A Lexical Analyzer Generator for JAVA) [18], developed by C. Scott Ananian, is applied to implement the morphology analysis. After completing the morphology analysis, the JLexper forms a transitional process for all words to their original forms, using a stemming algorithm. Next, it provides weights by measuring the frequencies of all words presented in the individual's documents through calculating term frequencies (TFs) and inverted document frequencies (IDFs).

The data learned by using the Bayesian Theorem can be used as the base for evaluating whether a specific web page has certain related information for the given domain. Bayesian Theorem can be simply defined as Eq. (1),

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)} \propto L(X \mid Y)P(X) \tag{1}$$

where $P(X)$ for random variables X and Y is a prior probability or a boundary probability. The term prior means that it does not consider any information for the event of Y. $P(X|Y)$ is a conditional probability of X, as Y is given. It is a posterior probability because it is determined by a specific value of Y. Also, $P(Y|X)$ is a conditional probability of Y, as X is given. $L(X|Y)$ is the likelihood of X, as Y is fixed. In this case, $P(Y|X) = L(X|Y)$.

The maximum probability can be calculated using a maximum posterior probability (MAP) in which the probability is to find a variable set of h that shows the highest possibility, as X is generated and can be obtained using Eq. (2). If $P(h)$ is assumed to be a constant value, the maximum likelihood (ML) can be replaced by the posterior probability instead of the maximum posterior probability. As the number of elements in a hypothesis set of H is presented as $|H|$, it can be assumed as Eq. (3). By using Eq. (3), h_{ML} can be determined as Eq. (4), as the ML probability that maximizes $P(X|h)$ is h_{ML} because the value of $P(h)$ in $P(X|h)P(h)$ becomes a constant of p.

$$h_{MAP} = \arg_{h\in H} \max P(h \mid X) = \arg_h \max \frac{p(X \mid h)P(h)}{p(X)} = \arg_{h\in H} \max P(X \mid h)P(h) \tag{2}$$

$$\forall h(\in H)P(h) = p = \frac{1}{|H|} \tag{3}$$

$$h_{ML} = \arg_{h\in H} \max P(X \mid h) \tag{4}$$

Sub-subheadings are to be in bold font. They should be indented by two characters and run in at the beginning of the paragraph. The start of this paragraph illustrates a sub-subheading.

In the searched documents, the words related to the given domain are not presented with the same frequency. Thus, as words related to the given domain appear more frequently, the higher the possibility that users must search the documents. One simple method to compute a weight is Term Frequency and Inverse Document Frequency (TF-IDF). The calculation of the weight by TF-IDF can be determined as Eq. (5), where $n_{i,j}$ is the number of frequencies of the word of t_i in document d_j,

and the dominator shows the number of frequencies of all words in document d_j. The weight of TF-IDF can be calculated by (TF × IDF).

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,i}}$$
$$idf_i = \log \frac{|D|}{|\{d_i : t_i \in d_i\}|} \quad (5)$$
$$tfidf_{i,j} = tf_{i,j} \times idf_i$$

In the preprocessing step, an ontology data dictionary is built. Ontology is a type of metadata in a specific domain, presented through the use of semantics in which information resources can be analyzed by a machine [19]. For example, the ontology describes the relationship between living things that can be classified as "species-genus-family-order-class-phylum-kingdom" using a formal lexicon. Such ontology can be used as a way that makes it possible to perform consistent communication between human beings and heterogeneous information systems. An ontology data dictionary is constructed with words related to a domain in order to recognize homonyms and synonyms, defining the relationships between words obtained through pre-learning. In the information integration process, the data dictionary filters words with synonyms in a reliability evaluation step and becomes a basis to assign name entities to terms of the same category.

3.2 Information Extraction

Information extraction is a process that selects a candidate group from the web pages searched by a given domain through verifying whether the searched pages are the related pages. Figure 1 shows the flow of the information extraction module. The information extraction module can be summarized in four steps: 1) data collection of web pages, 2) data extraction from the collected web pages, 3) verification of the inclusion of learned features, and 4) web page indexing.

For a given keyword in a domain, a crawler verifies the categories of the given keywords and extracts web page addresses through circulating the Web pages where the categories are classified by the pre-indexed Web pages or tags. Crawler plays a role to send the address of imported web pages to Wrapper by sequentially circulating web pages [20]. Wrapper is a tool that extracts and corrects information based on a type of rule-based system. It can be used to extract and correct the information of semi-structured data and structured data [21].

Wrapper extracts all data presented in web pages based on the addresses collected in the Crawler. The information of the web pages obtained by the Wrapper is to be taken through a feature test process that verifies whether it includes the learned features in the preprocessing step. The documents that include the learned features are indexed in order to perform the evaluation of its reliability and to easily configure metadata.

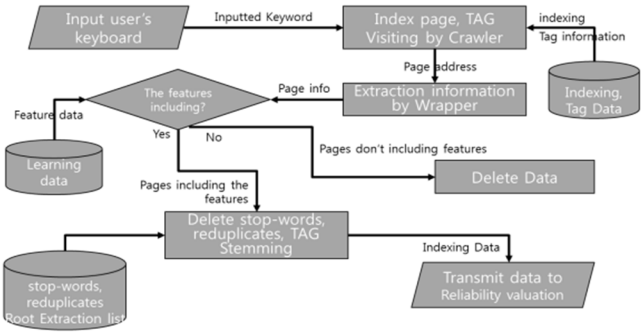


Fig. 1. Flow of the information extraction module

3.3 Reliability Evaluation

In the reliability evaluation step, data transmitted from the information extraction module is verified whether the data are reasonable for the given keywords. The reliability verification is performed by dividing it into two modules, such as category classification and reliability calculation modules. Figure 2 represents the flow of the reliability evaluation module. The data transmitted from the information extraction module are categorized based on the ontology data dictionary, established in the preprocessing step. Since each item has its own significance and weight values, it will be difficult to verify exact reliabilities if the categories for each item are not classified. For each item, categories search for terms that agree to the items and transmit the terms to the reliability calculation module. The reliability of the web page information in which the category classification is completed can be evaluated using the Naive Bayesian method.

$$\begin{aligned} C_{map} &= Category(w_1, w_2, ..., w_n) \\ &= \arg_c \max P(c \mid w_1, w_2, ..., w_n) \\ &= \arg_c \max \frac{P(c) \times P(w_1, w_2, ..., w_n \mid c)}{p(w_1, w_2, ..., w_n)} \\ &= \arg_c \max(c) \times P(w_1, w_2, ..., w_n \mid c) \\ &= \arg_c \max P(c) \times P(w_1 \mid c) \times P(w_2 \mid c) \times ... \times P(w_n \mid c) \\ &= \arg_c \max P(c) \times \prod_{t=1}^{n} P(w_1 \mid c) \end{aligned} \quad (6)$$

In the reliability calculation module, the reliability of the web page data, transmitted from the information extraction step, is evaluated using Naive Bayesian method based on the pre-learned probability, likelihood, and weight performed in the preprocessing step. Naive Bayesian method can be defined as Eq. (6) [22], whereas a random variable and a term set included in the literature are assumed to be c, and ω_n. $P(c)$ is the prior probability for occurring c, and $P(\omega_n|c)$ is the conditional probability for c. As the term of c is selected as the most exact category, C_{map} shows that ω_n contributes to the decision for how high the reliability is applied.

After evaluating the reliability, documents with over 80% reliability examined keyword matching to evaluate whether it contains data corresponding to the given keywords. On the other hand, documents with less than 80% reliability which does not include the given keyword are deleted from the memory.

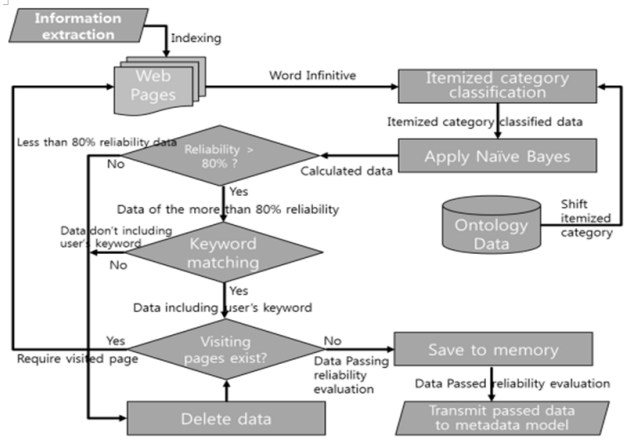


Fig. 2. Flow of the reliability evaluation module

The task is repeated depending on the number of web page progressed, until the web page addresses the obtained from Crawler is no longer available. If the extracted information is determined to be significant, stored data in the memory transmits as a construct part of information integration.

3.4 Information Integration

The information integration module is a critical part in the proposed method. Figure 3 shows the flow of the integration module. This module provides the final search results to users in a more convenient and easily readable ways. Unique named entities are given to the documents selected by the reliability evaluation module, with a reliability of more than 80% according to their categories. For each item, documents are processed in a named entity determination process through matching it with the terms defined in the ontology dictionary. This process is performed in order to integrate the terms that have related semantics, although they are in different formats extracted from various web pages. Tags are given to the documents for configuring the named entities to XML forms through the Wrapper. In the configuration module of the metadata, Wrapper performs a role to modify information than a role to extract information.

If the value of each items contain string form, by comparing strings of other items in the same category, words with the same meaning are deleted. Also, we use Java Wildcard String Matching Algorithm that integrates strings of one by selecting a word which is not included in the standard string. The items in numerical forms are examined whether it contains unit. Then, the units will be deleted if the category includes certain units. Although data has a value in numerical forms and because the value of numerical forms transform to string forms through the process of information extract, the unit has been deleted and thus, the value of numerical form is transformed into the INT form. The values transformed into the INT form suggest integration values which is the biggest number by comparing the differences of the items in the same category. In the case of the system with a specific purpose, the deleted unit can decide to belong to any unit in any item because the category is fixed. Therefore, integrated values are given in each category through Wrapper, as well as attached corresponding units in the predefined category. Data with unique named entities for each item is integrated and is finally stored as a form of files.

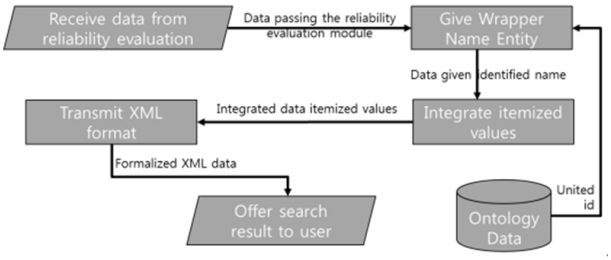


Fig. 3. Flow of the metadata construction module

Based on the keywords given by users, the XML metadata is stored as files, and the files are provided to users by outputting them as a type of XML document that has excellent readability.

4. EXPERIMENTAL RESULTS

In this study, a wheelchair information integration system for disabled people was implemented for verifying the efficiency of the proposed system [28]. According to the system configuration presented in Section 3, learning data was utilized for evaluating whether the obtained web pages have the related information with the user input keyword. In this experiment, 40 cases for both web pages with and without wheelchair data, respectively, were selected as training data. The data were used for preprocessing and the learning processes. In this paper, we removed the tags and stop words of all data to progress learning, and avoided learning of unnecessary information. Table 1 shows the features obtained from the learning process. Table 1(a) displays extracted features from web pages that include related data with wheelchair in its training data, while (b) shows extracted features from web pages that do not include wheelchair information. Although Table 1(b) presents information on wheelchairs, these are not the pages that the users want. This implies that learning for web pages without wheelchair information plays a role in screening the web pages recognized with wrong data in large documents, obtained by using integrated searches. As shown in Table 1, 14 items were selected for the features related to wheelchairs, and an order was determined for each feature item through analyzing the total number of appearances, average number of documents, and weights.

Table 2 represents the ontology data dictionary established by the feature items obtained from the learning process. Also, Table 2 demonstrates the feature of ontology data dictionary by classifying combinable terms, number of synonyms, synonyms, and unique category names. After completing the preprocessing, the entire process of extracting information is started as users input keywords. The Crawler collects web page addresses and the Wrapper extracts web page information based on these collected addresses. Wrapper, which received address of web pages from Crawler, extracted information of web pages as well.

The information of web pages refers to the sources of web pages. The extracted information from Wrapper is examined to see whether the information has the features obtained through learning in preprocessing. If no feature is available, web page information is deleted.

Table. 1 Features of Training Data for top 14 items

Feature Name	Total number of emergence	Document average	Weight	Ranking
Weight	73	1.825	0.125	1
Seat	66	1.65	0.113	2
Overall	63	1.575	0.108	3
Battery	58	1.45	0.099	4
Wheels	51	1.275	0.087	5
Width	50	1.25	0.086	6
Height	46	1.15	0.079	7
Length	38	0.95	0.065	8
Speed	38	0.95	0.065	9
Capacity	29	0.725	0.050	10
Clearance	26	0.65	0.044	11
Climb	20	0.5	0.034	12
Radius	17	0.425	0.029	13
HCPCS	9	0.225	0.015	14

(a) Features extracted for web pages with wheelchair data

Feature Name	Total number of emergence	Document average	Weight	Ranking
Weight	4	0.1	0.048	7
Seat	10	0.25	0.12	4
Overall	0	0	0	14
Battery	0	0	0	14
Wheels	17	0.425	0.204	1
Width	13	0.325	0.156	2
Height	13	0.325	0.156	2
Length	10	0.25	0.120	4
Speed	4	0.1	0.048	7
Capacity	2	0.05	0.024	9
Clearance	0	0	0	14
Climb	0	0	0	14
Radius	8	0.2	0.096	6
HCPCS	2	0.05	0.024	9

(b) Features extracted from non wheelchair web pages

Among the features, Width, Height, and Length were excluded in our examination because they are frequently used words by feature of tag. The evaluation module divides the corrected web page information into proper categories according to items which are configured based on the category in the ontology data dictionary established in the preprocessing step. The reliability of the web page information, after completing the categorization, is evaluated using the Naive Bayesian method. In the reliability evaluation, web pages information that obtained reliability of over 80% moves the keyword matching phase, and web pages information of less than 80% are discarded. For instance, if a user inputs the model name of "HS-1000," documents whose reliabilities are over 80% are moved to the integration construction module by determining the unique Name Entity by each category in order to provide integrated information with metadata configuration. For integrating information with different terms, the items, which are defined in the ontology dictionary, are to be searched, and a process that determines proper named entities for each item is performed. Figure 4 shows the ontology data dictionary which is constructed by Protégé [23].

Table. 2 Feature ontology data dictionary

Feature Name	Possible combination of the words	Number of synonyms	High freq. synonym	Person-specific category
Weight	Overall, Battery, Capacity	1	Heaviness	Overall_Weight Battery_Weight Weight_Capacity
Seat	Width	1	Chair	Seat_Width
Overall	Length, Width, Height	3	Total	Overall_Length Overall_width Overall_Height
Battery	Weight	0		Battery_Weight
Wheels		1	Tire	Wheels_Size
Width	Overall, Seat	2	Breadth	Overall_Width Seat_Width
Height	Overall	1	Tallness	Overall_Height
Length	Overall	1	Distance	Overall_Length
Speed		1	Rate	MAX_Speed
Capacity	Weight	2	Capability	Weight_Capacity
Clearance		0		Ground_Clearance
Climb		2	Incline	Grade_Climb
Radius		0		Turn_Radius
HCPCS		0		HCPCS
Front	Wheels	0		Front_Wheel
Rear	Wheels	0		Rear_Wheel
Suspension		0		Suspension
Brake		0		Break
Mortor		0		Mortor_Type
Size	Wheels, Battery	0		Wheels_Size Battery_Size
Traver		1	Maximum	Traver_Range
Charger	Battery	0		Battery_Charger

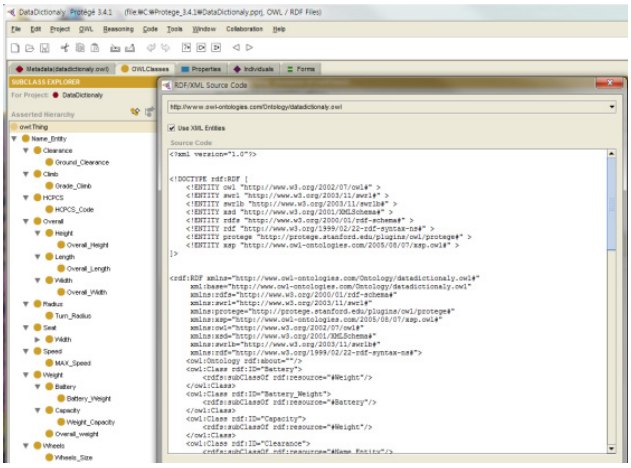


Fig. 4. Ontology data dictionary for determining named entities

The next step is to apply tags for configuring the named entities by the Wrapper to actual XML formats. Figure 5 shows the results of the application of tags for the wheelchair "HS-1000." Finally, XML metadata can be produced by matching it with the keywords input by users. Finally, the results are provided to users as a type of document that has an excellent readability in grid tables. The right side page in Fig. 5 shows the final document form for users.

Table 3 presents the results of the evaluation Set for wheelchair web pages. Experiments were performed by selecting 20 web pages with and without wheelchair information, respectively. The reliability for each item was more than 80%, and the average reliability in the overall system was 82.595%. In particular, in the feature test, all 20 web pages with wheelchair information were well selected, and 16 web pages were deleted from the 20 web pages without wheelchair information in order to decrease the calculation complexity, thus leading to an improvement in the system performance.

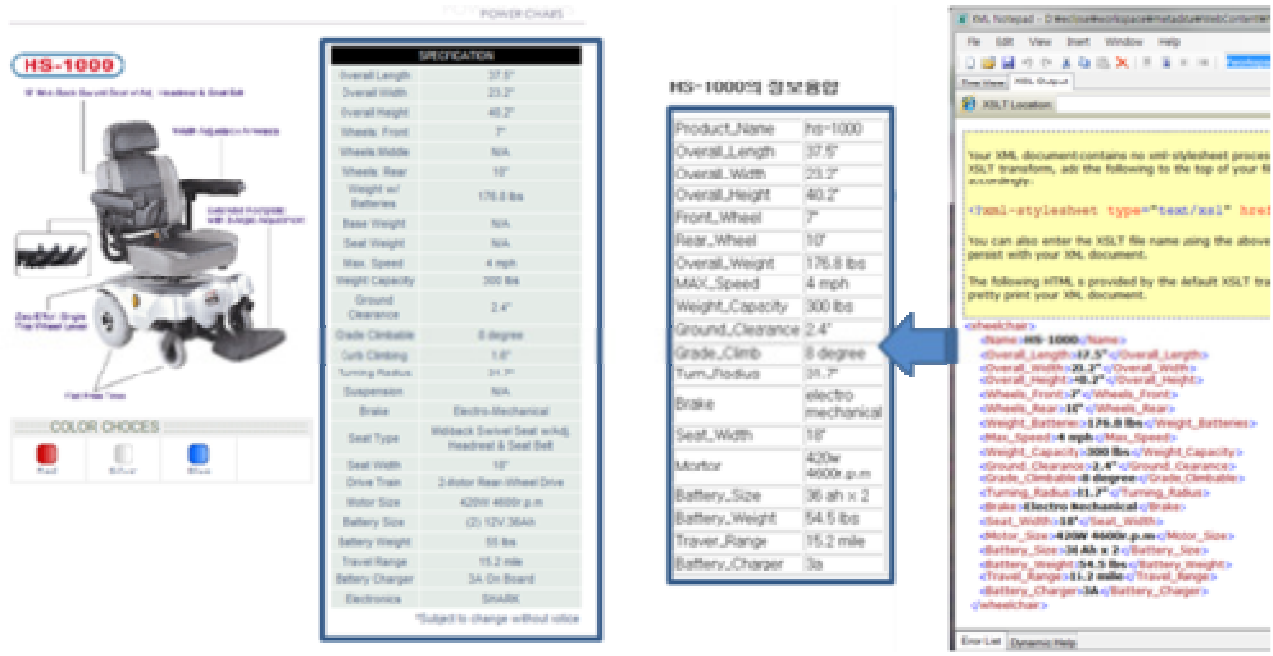


Fig. 5. Named entities by the Wrapper and their integration with XML metadata

Table. 3 Evaluation results

	Web pages with wheelchair information		Web pages without wheelchair information	
Feature Test	20 Web pages: 100%		4 Web pages: 20%	
Category Classification for Each Item	Total 292 times: 94.8% Success: 277 times Failure: 15 times		Total 23 times: 91.3% Success: 21 times Failure: 2 times	
Reliability Evaluation	Total 20 times: 95% Success: 19 times Failure: 1 times		Total 20 times: 100% Success: 0 times Failure: 20 times	
Applying Wrapper Named Entities	Total 292 times: 81.8% Success: 239 times Failure: 53 times		Total 23 times: 82.6% Success: 19 times Failure: 4 times	
Value Integration for Each Item	Total 292 times: 88.98%		Total 23 times: 76.21%	
	Character type: 94 times Success: 71 Failure: 23 75.5%	Numerical type: 198 times Success: 172 Failure: 26 86.8%	Character type: 9 times Success: 7 Failure: 2 77.7%	Numerical type: 14 times Success: 12 Failure: 2 85.7%
System Average	82.595%			

The performance of this system was compared that of other studies. The subjects used in this comparison have similar system configurations to the proposed system and represent similar results for each module even though the methods applied in these studies are different.

The comparison was performed in an indirection manner because of the limited number of comparable studies with the same methods as this study. In addition, most of the other studies selected only a single module. Table 4 presents the results of the comparison and evaluation of the proposed system.

Table. 4 Performance comparison and evaluation

Items	Category Classification	Integration for Each Item	Value Integration for Each Item	Overall System
Choi & Park [24]	90.22%	84.32%	41.49%	72.01%
Corcho et al.[9]	92.27%	84.44%	53.68%	76.79%
Kim et al.[25]	54.37%	86.30%	84.12%	74.93%
Lee et al.[26]	84.21%	76.44%	80.29%	80.31%
Kim et al[11]	69.29%	80.76%	77.74%	75.93%
Assali & Zanghi [10]	68.73%	81.29%	79.42%	76.48%
Guha & Mccool [27]	72.52%	76.15%	63.29%	70.65%
This Study	93.05%	82.2%	86.25%	87.16%

In Table 4, the category classification module classifies the related features extracted from web pages into a single category, and the integration module for each item is to integrate similar terms by identifying their homonyms and synonyms. The value integration for each item integrates the items that have numerical type values. The result of category classification shows higher classification rate of 2.83% than [24], which has the best performance between existing researches. The integration of values by items exhibits higher performance result of 2.13% than [14]. The integration module by items indicates a lower result of 4.1% than [11], which has the highest integration rate among existing researches. However, the retrieval accuracy of the overall system shows higher accuracy of 87.16%, improving of 6.85%, when compared with [10], which shows the highest accuracy between existing retrieval systems.

5. CONCLUSION

In this study, an information integration system that provides more qualified information located on web pages was proposed. The proposed system integrates scattered heterogeneous data into a single document following the sequence of preprocessing, information extraction, reliability evaluation, and information integration. In addition, a wheelchair information integration system for disabled people was implemented for verifying the efficiency of the proposed method. The contribution in the proposed method is as follows. Firstly, a classification process for each item, including their homonyms and synonyms by establishing an ontology data dictionary, is performed. The problem caused by semantic relatedness through applying unique named entities is solved. Secondly, integrated information corresponding to the keywords input by users, instead of addressing links as users input keywords, is performed. Thirdly, we used the database in only the learning of preprocessing in order to minimize the use of database, which secured the scalability by learning with ontology and data of web pages in difference domains.

This method can provide the usability to users that are not accustomed to information retrieval. In addition, it can solve problems of inconvenience to experience in order to obtain the information which the user wants, thus, reducing retrieval time. This method will improve certain inconveniences in obtaining information and prevent wasted time; it provides exact and highly readable search results. In future studies, an extension of this system for providing the integrated information of various multimedia data, including images and movies, will be proposed.

ACKNOWLEDGEMENTS

This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2012-H0301-12-3005). This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST)(2012-047759).

REFERENCES

[1] T.B. Lee, J. Hendler, and O. Lassila, "The Semantic Web," Scientific American, May 2001.

[2] J. Martin, *Strategic Data Planning Methodologies*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1982, p.127.

[3] J.C. Song, D.I. Lee, and B.J. Moon, "Standardization of Semantic Web and Development Trends of Technical Factors," National IT Industry Promotion Agency, [IITA] Weekly Technological Trends, 2002.

[4] S.Y. Park, "Comparative Evaluation of Directory Services Provided by Major Korean Search Portals: In the Field of Computer and Internet," Korean Society for Library and Information Science, Journal of the Korean Library and Information Science Society, 2009, pp.215-234.

[5] A. Gomez-Perez, M. Fernandez-Lopez, and O. Corcho, "Ontological Engineering: With Examples from the Areas of Knowledge Management," E-commerce and the Semantic Web. Springer, 2004.

[6] S.E. Shin and Y.H. Seo, "Semantic-based Query Generation For Information Retrieval," International Journal of Contents, vol.1, no.2, 2005, pp.39-43.

[7] http://www.Nate.com

[8] M. Y. Ra, "A Metadata-based Data Grid System for the Integration of Military Information," Korea Society of Computer Information Papers, vol.13, no.2, 2008, pp.95-103.

[9] O. Corcho, P. Alper, P. Missier, S. Bechhofer, and C. Goble, "Grid metadata management: Requirements and architecture," *Proc. 8th IEEE/ACM International Conference on Grid Computing*, 2007, pp.97-104.

[10] A. A. Assali and H. Zanghi, "Automated Metadata Hierarchy Derivation," Information and Communication Technologies, vol.1, 2006, pp.505-510.

[11] D. K. Kim, K. J. Jeong, H.S. Shin, and S.T. Hwang, "An XML Schema-based Semantic Data Integration," Korean Institute of Information Scientists and Engineers Papers: System and Theory, vol.33, no.9, 2006, pp.563-573.

[12] http://developer.yahoo.com/searchmonkey/

[13] J.H. Kim, H.Y. Kwak, and H. Kwon, "RDB Schema Model of XML Document for Storage Capacity and Searching Efficiency," Journal of Korea Contents, vol.6, no.4, 2006, pp.19-28.

[14] RDF, http://www.w3.org/RDF

[15] W. H. Yu and H. J. Koh, "Design of a RDF Metadata System for the Searching of Application Programs," Journal of Korea Contents, vol. 5, no. 6, 2005, pp.1-9.

[16] OWL, http://www.w3.org/TR/owl-features

[17] D. Braverman, "Learning filters for optimum pattern recognition cognition Learning filters for optimum pattern recognition," Knowledge Acquisition, 1993, pp.280-285.

[18] http://www.cs.princeton.edu/~appel/modern/java/JLex/

[19] T. Gruber, "A translation approach to portable ontology specifications," Knowledge Acquisition, vol.5, 1993, pp.199-220.

[20] http://tartarus.org/~martin/PorterStemmer/

[21] J. H. Cho and H. G. Molina, "Parallel Crawlers," Technical Report, Stanford University, 2001.

[22] http://johannburkard.de/blog/programming/java/java-wild card -string-matching.html

[23] Protégé, http://protege.stanford.edu/

[24] J. H. Choi and Y. T. Park, "Ontology-based Automated Metadata Generation Considering Semantic Ambiguity," Journal of KISS-Software and Applications, vol.33, no.11, 2006, pp.986-998.

[25] S.S. Kim, S.H. Myaeng, and J.M. Yoo, "A Hybrid Information Retrieval Model Using Metadata and Text," Korean Institute of Information Scientists and Engineers Papers : Database, vol.34, no.3, 2007, pp.232-243.

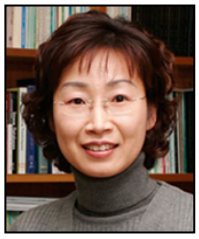
[26] I.K. Lee, D.S. Hwang, S.T. Seo, and S.H. Kwon, "Ontology Integration based on Meta Ontology," Korean

Institute of Intelligent Systems Papers, vol.34, no.3, 2007, pp.604-613.

[27] R. Guha and R. McCool, "Tap: Towards a Web of Data," http://tap.stanford.edu/.

[28] K.-Y. Kim, Y. S. Kim, M. R. Schmeler, "Remote Decision Support for Wheeled Mobility and Seating Devices," Expert Systems with Applications, Vol. 39, No. 8, 2012, pp. 7345–7354.

Hyung-Wook Shin
He received the B.S., M.S in computer science from Gwangju university and Chonnam university, Korea in 2008, 2010, respectively. His main research interests include Semantic Web Information Integration. Present, He is working on development 2 team of Inzisoft company.



Hyung-Jeong Yang
She received her B.S., M.S. and Ph. D from Chonbuk National University, Korea. She is currently an associate professor at Dept. of Electronics and Computer Engineering, Chonnam National Uni versity, Gwangju, Korea. Her main research interests include multimedia data mining, pattern recognition, artificial intelligence, e-Learning, and e-Design.

Soo-Hyung Kim
He received his B.S. at Dept. of Computer Engineering, Seoul National University, and M.S. and Ph.D. at Dept. of Computer Science, Korea Advanced Institute of Science and Technology, Korea. He is currently a professor at Dept. of Electronics and Computer Engineering and a vice-Dean of the Engineering College, Chonnam National University, Gwangju, Korea.

Guee-Sang Lee
He received his BS in electrical engineering and his MS in computer engineering from Seoul National University, Seoul, Rep. of Korea, in 1980 and 1982, respectively. He received his PhD in computer science from Pennsylvania State University, University Park, PA, USA, in 1991. He is currently a professor of the Department of Electronics and Computer Engineering at Chonnam National University, Gwangju, Rep. of Korea. His main research interests are image processing, computer vision, and video technology.

Kyoung-Yun Kim
He received his B.S. and M.S. from Chonbuk National University, Korea and Ph. D from the University of Pittsburgh, USA. He is currently an associate professor at Dept. of Industrial and Systems Engineering at Wayne State University, Detroit, USA. His main research interests include Design Informatics, CAD/CAM and Telerehabilitation.

Sun-Hee Kim
She received the B.S in Multimedia from Korean Educational Development Institute in 2004 and the M.S. degree in Computer Science from Dongguk University, Korea in 2006. She received the Ph. D. degrees in Computer Science from Chonnam National in 2011. She recently works in Carnegie Mellon University as a researcher. Her research interests include Data Mining, Sensor Mining and Bioinformatic.

Do Luu Ngoc
He received the B.S. in computer science from Chonnam National university, Korea in 2012, and also he is currently a student for M.S at Dept. Electronics and Computer Engineering from Chonnam National university. His main research interests include Mobile Application, SemanticWeb, and Information Integration.