Received 27 October 2005
Revised 6 February 2006
Accepted 8 February 2006

Developing and evaluating an IT specification extraction system

Chyan Yang
Institute of Business and Management & Institute of Information Management, National Chiao Tung University, Taipei, Taiwan

Liang-Chu Chen
Institute of Information Management, National Chiao Tung University, Hsinchu, Taiwan, and

Chun-Yen Peng
Taiwan Semiconductor Manufacturing Company, Hsinchu, Taiwan

Abstract

Purpose – This paper seeks to establish an extraction system for an information technology (IT) product specification named ITSIES which combines the natural language process (NLP) with the ontology concept and also to evaluate the system's effectiveness in advance.

Design/methodology/approach – The development of the system is based on a prototype design and performance validation. This study adopts four classes of IT specification (PC, Unix server, Monitor, and Printer) that follow IBM's and HP's product lines as the baseline information in order to construct the extraction system in GATE (General Architecture for Text Engineering) tools and to examine the IT product specification with other brands and patterns. Additionally indices are adopted such as precision, recall, and F-measure as the matrices for evaluating system performance.

Findings – The performance shows that the average recall, precision, and F-measure are all over 90 per cent, revealing that the JAPE (Java Annotation Patterns Engine) grammar rules in the IT domain are reasonably good and generally in line with expectations.

Originality/value – The paper proposes an integrative framework to examine IT product specification information and demonstrates that the system is effective for IT application.

Keywords Knowledge management, Communication technologies

Paper type Research paper

Introduction

With the arrival of the dynamic environment and the knowledge economy, knowledge has been treated as one of the most important assets that can enhance competitive advantages. For a company to be a leader among its competitors, it is important to ensure that the best corporate knowledge must be available and applied to employees and customers. Thus, how to explore and exploit knowledge so as to maintain an enterprise's high competitiveness is a critical task for knowledge workers. In general, the knowledge base or expert system is an effective methodology when a firm implements knowledge management (Liebowitz, 2001). The knowledge base, a repository for collecting, capturing, identifying, classifying, and organizing a firm's knowledge, is in fact the knowledge library of a company. The activities of knowledge workers rely heavily on document control centers in all companies. The artifacts of document control centers are manifested as an electronic library in an on-line networked computer system. Furthermore, the knowledge source of document control centers is usually the world wide web (www).

The Electronic Library
Vol. 24 No. 6, 2006
pp. 832-846

0264-0473
DOI 10.1108/02640470610714251

The www, an enormous repository of information, is rapidly becoming the main knowledge thesaurus for many industries. Most of the information therein is presented by unstructured documents in natural languages and dispersed on different Web sites. A search engine usually needs much time and effort to select related documents and it annotates these documents manually when extracting information. Therefore, an efficient approach has to be developed for collecting and converting natural language documents into a readable computer format for knowledge workers' needs, which is indeed a complex task.

Information extraction (IE) is an important front-end technique for knowledge discovery, data mining, and natural language interface to databases (Jung *et al.*, 2005). The major concern of IE is to build effective systems or rules that address relevant information while ignoring extraneous and irrelevant information (Cowie and Lehnert, 1996; Jung *et al.*, 2005). Information extraction from the www is usually performed by a software module named wrapper that is proposed by many scholars (Embley *et al.*, 1999; Soderland, 1999). A critical web technology and application challenge in IE is to acquire domain portability (Jung *et al.*, 2005). To extract specification information from Web pages, an information extraction pattern for specific domain knowledge has to be developed – that is, each application of information extraction needs a separate set of rules mapped to the specific domain and presentation style.

For most companies and individuals, collecting product specifications for comparison and evaluation is a necessary task before purchasing something. Specifications can be viewed as semi-structured documents with standard terminologies (Thirunarayan *et al.*, 2005), or more specifically, information technology (IT) products. Despite product specifications having a common format, a little difference still exists in various Web sites. In general, two reasons are revealed herein that explain why it is hard to obtain valid IT product information. Firstly, the life cycle of an IT product is very short, as its specifications always change with a new product release. Next, this kind of information usually disperses itself in many different brands and formats which exist on different Web sites. For example, the presentation of CPU cache information in IBM shows up as "Level1 cache: 4MB", but in HP it shows up as "4MB L1 cache".

Recent advances in natural language processing (NLP) provide a solution to perform document annotation and information extraction tasks. Through customized and pre-defined process flows, NLP tools can extract information accurately from Web pages in a specific domain (Khelif and Dieng-Kuntz, 2004). On the other hand, the ontology is a description of the concepts and relationships and provides a new model about information presentation. Embley *et al.* (1999) claimed that an ontological approach can find and classify Web pages of interest for a given application, which is an important development in information extraction. In general, ontology has been used effectively to support domain knowledge extraction (Alani *et al.*, 2003) and knowledge sharing (Edgington *et al.*, 2004), but comparatively few research studies have been conducted on the IT product perspective for integrating NLP and the concept of ontology.

The objectives of this study are presented as the following:

- to build up a prototype system for extracting IT product specifications automatically from Web pages efficiently and accurately;

- to save the extracted information into the structure of ontology language to offer wide applications in information seeking; and
- to compare the system's effectiveness with the extraction performance and to apply the same rules to other IT brands and patterns.

Related work

Much research effort has been focused on enhancing information retrieval and extraction schemes to meet the needs of knowledge workers. For example, Strzalkowski and Vauthey (1992) tried to build up an information retrieval system which uses advanced natural language processing techniques to enhance the effectiveness of traditional key-word based document retrieval. Alani *et al.* (2003) took up the Artequakt project, which seeks to automatically extract knowledge about artists from the web, populate a knowledge base, and generate personalized narrative biographies. Parts of some studies have concentrated on exploring the process of information extraction. For instance, Liu *et al.* (2001) claimed that the information extraction process involves three steps:

(1) identifying regions of interest on a page;

(2) identifying semantic tokens of interest on a page; and

(3) determining the nesting hierarchy for the content presentation of a page.

The outcomes in each step include a set of rules to generate a wrapper program code.

Some researchers have also put attention towards integrating various technologies. Thompson's research team developed an IE system that applies the method of active learning to reduce annotation effort from natural-language documents (Thompson *et al.*, 1999). Embley *et al.* (2005) proposed a solution based on document-independent ontology for extracting target information from HTML tables with an unknown structure. Sung and Chang (2004) used inductive learning to design an investigation agent system for extracting business information from Web pages. For automatically reorganizing and summarizing specification content, Thirunarayan *et al.* (2005) described a coarse-gain technique to extract alloy specification information. To summarize, rarely do studies focus on domain knowledge about IT products by using the information extraction approach. It is this reason why the paper herein studies such a theme.

System architecture

IT products have diverse and complicated specifications in that it is difficult to develop common patterns or standards for all IT products. This study adopts a part of IT products to develop a prototype system. The IT product specification information extraction system (ITSIES), which is segmented into four main applications (Personal Computer, Unix Server, Monitor, and Printer), is developed under the Microsoft platform. Each application, which has a related language resource (e.g. web corpus, ontology) and a process resource (e.g. tokeniser), is an independent sub-system that extracts information from the IT product specification knowledge domain.

We partition each application into three essential processes – annotation, JAPE transduction (Java Annotation Patterns Engine), and DAML + OIL exporter (DARPA Agent Markup Language + Ontology Interchange Language). The system architecture and the workflow in ITSIES are shown in Figure 1. Based on the

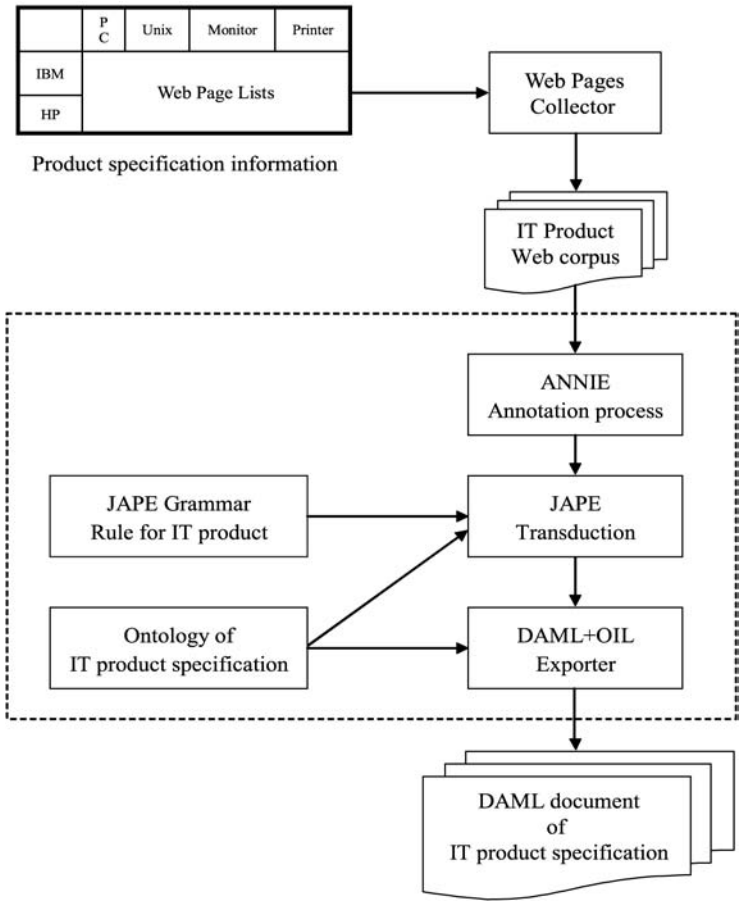

Figure 1.
System architecture and workflow of IT product specification

pre-defined web list, the system collects documents and saves them into specific paths as a corpus from different IT products. Next, the NLP tools load the corpuses to annotate these documents, and then the JAPE transducer loads the JAPE grammar rule to mark up the product specification. Finally, the DAML + OIL exporter generates an output file of specifications in DAML document format.

Web page collection

This study adopts Teleport Pro, a widely-used offline browsing software, as the data gathering tool. For enhancing the precision of the system, we take the IT products of HP and IBM as best practices in order to develop and examine the system. IBM and HP are famous international companies in the computer field, and their product lines are richer and their specifications are more representative than other competitors. Therefore, we randomly select three to five products from HP and IBM per product line, for a total of 34 web pages of information which have been downloaded for designing the extraction system. Additionally, 22 IT products are collected from seven

companies to evaluate the effectiveness of the system. All of the sources of Web sites are shown in Table I.

Annotation process

GATE (General Architecture for Text Engineering) is a specialized tool of Java Bean that has been built over the past eight years by the Sheffield University NLP group. In general, there are three types of resources which are helpful for enhancing system functionality by using GATE components (Bontcheva *et al.*, 2003). First, language resources (LRs) aim at representing entities such as lexicons, corpora, or ontologies. Next, processing resources (PRs) depict those entities that are primarily algorithmic such as parsers, generators, or modelers. Lastly, visualization means that components participate in GUIs through visual resources (VRs).

GATE provides a set of reusable and extendable language processing components for common NLP tasks, known collectively as ANNIE (A Nearly New Information Extraction System). ANNIE relies on finite state algorithms and the Java Annotation Patterns Engine (JAPE) language to produce precision and recall figures for an entity recognition of roughly 90 per cent (Maynard and Cunningham, 2003). Therefore, this study adopts GATE as an annotating tool for achieving the annotation process effectively. After this processing, web corpora are separated into thousands of tokens. Each token has an attribute such as token type, length, and position, etc. Figure 2 shows the annotation results.

The major task of the ANNIE process in ITSIES is to separate a Web document into tokens and add affiliated attributes to these tokens (Bontcheva *et al.*, 2003). The default function of ANNIE only identifies data concerning "name", "address", and "date", etc. Therefore, to extract information from specific domain knowledge, customization is necessary. After considering the flexibility and integration with the ontology model, the JAPE is adopted.

JAPE transduction

JAPE provides a finite state transduction over annotations based on regular expressions. The JAPE grammar consists of a set of phrases, each of which covers a set of patterns or action rules. The JAPE rule is combined with LHS (left-hand-side) and RHS (right-hand-side). The LHS includes an annotation pattern that may contain

Table I.
The web sources of IT product specification

Company	Web site	IT products
Training		
IBM	www.ibm.com/	PC, Unix Server, Monitor, and Printer
HP	www.hp.com/	PC, Unix Server, Monitor, and Printer
Testing		
Acer	www.acer.com.tw/	Monitor
Asustek	tw.asus.com/	PC
BenQ	www.benq.com.tw/	Monitor
Canon	www.canon-asia.com/	Printer
Epson	http://w3.epson.com.tw/ett/	Printer
Sony	www.sony.com.tw/	PC
SUN	www.sun.com/	Unix Server

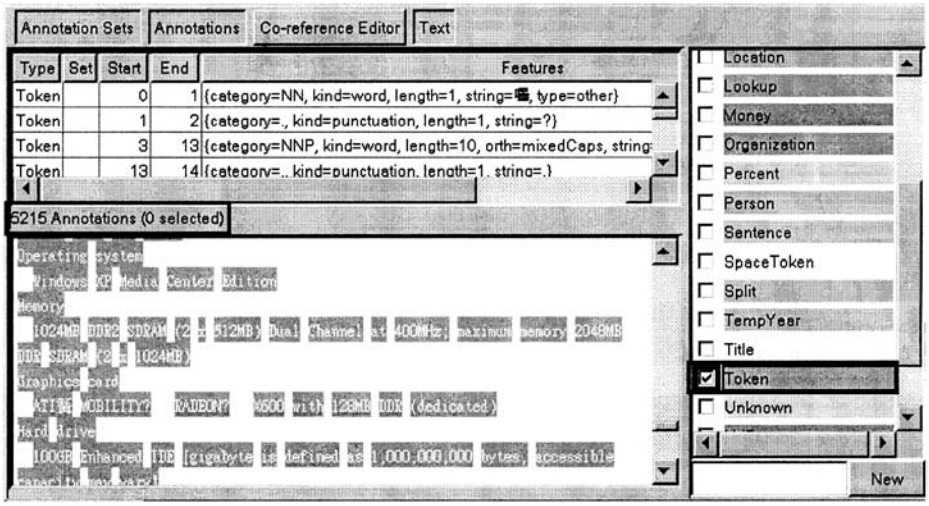

Figure 2.
Result of the annotation process

regular expression operators. The RHS consists of annotation manipulation statements. Annotations matched on the LHS of a rule may be referred to on the RHS by means of labels that are attached to pattern elements.

GATE supports ontology-aware grammar transduction, and this allows a JAPE transducer to match not only those features on the LHS exactly, but also to match any features that are subclasses of those specified in the JAPE rule. The JAPE grammar rule is very suitable for extracting an information entity when the entity has a specific text structure. For example, a monitor usually provides several different resolution modes, e.g. "1024 × 768 @ 60Hz" or "800 × 600 @70Hz". The mode has a common text structure as:

$$\{\text{number}\}\{\text{x}\}\{\text{number}\}\left(\{\,|\,\text{at}\,\}\right)\{\text{number}\}\{\text{Hz}\}$$

Based on grammar rules that we develop, the JAPE transducer tries to complement the information entity which may be composed by several tokens. Figure 3 shows the process screen whereby the JAPE transducer marks the information entity up when the text format is matched with the JAPE grammar rules.

Ontology and JAPE extraction rule

To enable the ontology-aware grammar feature, an ontology for IT product specification is developed. Based on the characteristic of the IT product, the specification information is depicted as the following:

- *Personal Computer:* Model Name, CPU Model, CPU Speed, Memory Type, Memory Size, Hard Disk Type, and Hard Disk Size.
- *Unix Server:* Server Model, CPU Model, CPU Speed, Cache Memory Size, Memory Type, Memory Size, Max. Memory Size, OS Version, and Hard Disk Type.
- *Monitor:* Model Name, Monitor Type, Monitor Size, Pixel Pitch, Resolution Mode, Max Resolution Mode, Recommend Resolution Mode, Refresh Rate, and Web Price.
- *Printer:* Model Name, Print Speed, Print Quality, and Memory Size.

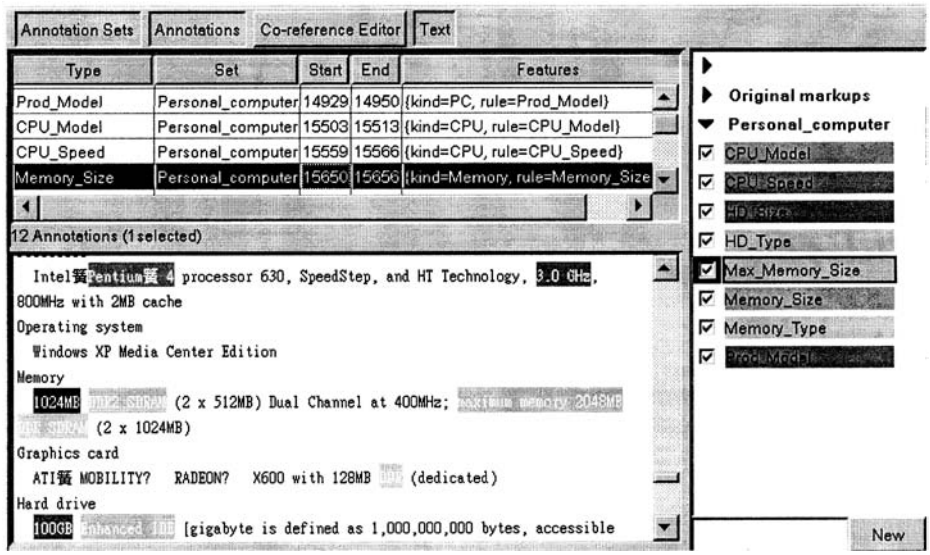

Figure 3.
Information entity marked up by the JAPE transducer

Figure 4 shows the ontology structure of a PC which describes the relation of information as an example. After considering the compatibility and precision, GATE built-in tools are operated in order to develop the ontology. Figure 5 shows the interface of the system.

GATE saves these ontologies of IT product specification as DAML files. These files have to be reloaded into GATE as a language resource, and the JAPE transducer and DAML exporter will use it to perform the extraction process.

DAML + OIL exporter

Ontology Interchange Language (OIL), which was developed in the OntoKnowledge project, permits semantic interoperability among Web resources. The syntax and

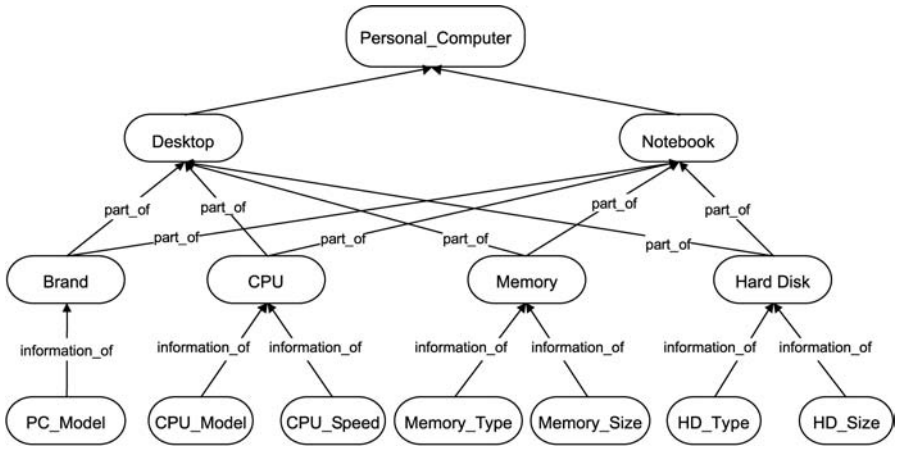

Figure 4.
Ontology structure for personal computer specification

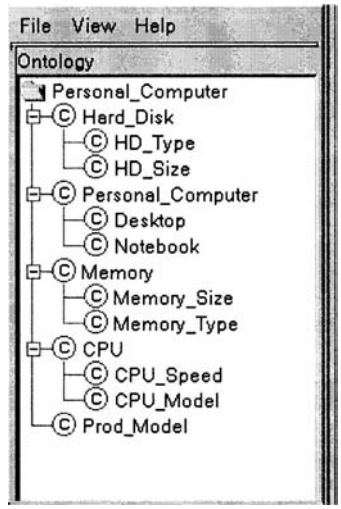

Figure 5.
Example of built-in ontology development

semantics of OIL provide modeling primitives commonly used in frame-based approaches to ontological engineering. The DARPA Agent Markup Language (DAML), a project sponsored by the US government, aims at providing the ground for the semantic web. The purpose of DAML is to develop languages, tools, and techniques to reflect more machine-understandable content on the Web. DAML + OIL, targeting the same objective as OIL, was developed by a joint committee from the US and the European Union (IST) in the context of DAML. The DAML + OIL Export is a GATE process resource that allows the information segment found in documents to be exported as instances of a specified ontology. When a corpus is processed with ANNIE and the JAPE transducer, GATE will mark up the information that one needs to extract. When the DAML + OIL exporter processes the corpus and for each information segment found that is of some type (e.g. CPU_type), if a corresponding concept with the same name as the information type (e.g. CPU_type) exists in the ontology, then a new DAML instance is generated in the export file.

Analysis and comparison

Evaluation procedures are often concerned mostly with cross-validation, or splitting in training and testing sets (De Sitter *et al.*, 2004). To assess the ITSIES performance, three phases are set up. Phase 1 is called manual annotation, which is a baseline of performance evaluation. The purpose of this phase is to annotate all selected Web documents manually to mark up all target information entities. Phase 2 aims at automatic annotation through the JAPE grammar rules. All applications of ITSIES are executed to mark up and extract information entities from selected Web pages by an automatic process. To verify the effectiveness of the developing system, phase 3 is performed to compare the information entities that are marked up manually and extract them by an automatic process to calculate the performance indices, including recall, precision, and F-measure. In addition, some new IT products as testing data are examined. Before performing the evaluation stages, some definitions are described below:

- *Information entity:* an IT product specification that may be formed from one or several lexical tokens. For example, "1024 x 768@75Mhz" is an information entity in monitor resolution mode.
- *Target information entity (TIE):* an IT product specification meets our information extraction scope and hides itself in a selected web document.
- *Information entity extracted (IEE):* a lexical construction that is extracted by means of the JAPE grammar rules.
- *Correct information entity extracted (CIE):* a correct IT product specification that is extracted by the JAPE grammar rules.

Experimental results and discussion

The goal of ITSIES is to extract the necessary information entities from IT firms' web pages. A total of thirty-four web pages are collected from the Web sites of IBM and HP for system training and tuning. Table II shows the extraction results of the tokens and information entities.

The product tokens of HP are in general larger than those for IBM except Unix servers. Additionally, the numbers of extracted information entities have significant differences which appear in the products of monitors and printers. For a monitor, IBM's information entities are higher than HP's. This is contrary to the printer, where the information entities in HP are greater. These findings present an interesting issue in the relationship between information tokens and information entities, which may be a future area to explore.

Although ITSIES can extract information entities successfully from Web pages, the effectiveness and efficiency of a system are more important tasks in performance evaluation. In general, the methodology of evaluation is either quantitative or qualitative (Navigli *et al.*, 2004). Quantitative evaluation measures the performance of the various software algorithms that constitute the extraction system. Qualitative evaluation assesses the adequacy of the information extraction method for a specific knowledge domain.

To measure the performance of ITSIES, three performance indices are considered in this study – precision, recall, and *F*-measure (Chung *et al.*, 2005; Lavelli *et al.*, 2004; Popov *et al.*, 2003). These indices focus on how well the system performs at identifying the relevant information. Precision indicates how many of the extracted information entities are correct and helps the user filter irrelevant results. The recall value specifies how many of the information entities that should have been found, and it helps the user discover all the relevant results. The *F*-measure is a summarized metric which combines recall and precision into a single value. The average performance is better when the *F*-measures are greater. Based on the preceding definitions and discussion, the formulae are stated below:

$$\text{Precision} = \text{Number of correct information entities extracted} / \text{Number of information entities extracted}$$

$$\text{Recall} = \text{Number of correct information entities extracted} / \text{Number of target information entities}$$

$$\text{F} - \text{measure} = (2*\text{Precision}*\text{Recall})/(\text{Precision} + \text{Recall})$$

Product catalogues	Web site	Product name	Total tokens	Extracted information entities
Personal computer	HP	Pavilion dv4030us Notebook	5190	11
		Pavilion zv6010us Notebook	4900	10
		Media Center zd8110us Notebook	5215	9
		Pavilion a1040n Desktop	5104	12
		Media Center m7070n Photosmart PC	5103	10
	IBM	ThinkPad T42 Notebook	2128	11
		ThinkPad X41Tablet	1806	10
		ThinkPad R50e Notebook	1812	10
		ThinkCentre S50 Desktop	1816	10
		ThinkCentre M51 Desktop	1851	10
Unix server	HP	RX8620	892	18
		RX7620	860	16
		RX4640	1079	25
		RP8420	1120	16
		RP7420	1230	15
	IBM	p5 590	1016	13
		p5 570	1080	18
		p5 550	1017	11
		p5 520	1031	11
		p5 510	772	11
Monitor	HP	vs1717 Flat-Panel LCD Monitor	4932	11
		f1905 19 LCD Flat-Panel Monitor	4928	11
		Pavilion mx704 17 Flat-Screen CRT Monitor	4903	11
		s7540 CRT Monitor	527	11
	IBM	ThinkVision L170p Monitor	1117	27
		ThinkVision L150 Monitor	1050	19
		E74M 17 inch CRT Monitor	1108	25
		E54 15" Monitor MPRII	809	10
Printer	HP	Inkjet 1200dn printer	3766	21
		DeskJet 3845 Printer	4905	6
		Color LaserJet 2600n Printer	3431	11
	IBM	Infoprint Color 1357 Printer	782	5
		Infoprint 1412 Printer	807	9
		Infoprint 1352 Printer	889	8

Note: Personal computer corpus (include desktop and notebook)

Table II.
Extracting results of tokens and entities in training IT products

All of the applications are performed and calculated with these performance indices. Table III shows the results.

The results show that the JAPE grammar rules have been proven well by high recall, precision, and *F*-measures. Our proposed system has an average recall rate of 100 per cent, an average precision rate of 91.3 per cent, and an average *F*-measure of 95.5 per cent for Personal Computer. As for other IT products, the recall rate averages to 95 per cent, the precision rate is from 91.7 per cent to 96.2 per cent, and the F-measure is from 93.2 per cent to 95.6 per cent. These findings reveal that the JAPE grammar rules in the IT domain are reasonably well and generally in line with our expectations.

Company	Product name	No. TIE	No. IEE	No. CIE	Recall (%)	Precision (%)	F-measure (%)
(a) Personal Computer							
HP	Pavilion dv4030us Notebook	10	11	10	100.0	90.9	95.2
	Pavilion zv6010us Notebook	10	10	10	100.0	100.0	100.0
	Media Center zd8110us Notebook	8	9	8	100.0	88.9	94.1
	Pavilion a1040n Desktop	10	12	10	100.0	83.3	90.9
	Media Center m7070n Photosmart PC	9	10	9	100.0	90.0	94.7
IBM	ThinkPad T42 Notebook	10	11	10	100.0	90.9	95.2
	ThinkPad X41Tablet	10	10	10	100.0	100.0	100.0
	ThinkPad R50e Notebook	9	10	9	100.0	90.0	94.7
	ThinkCentre S50 Desktop	9	10	9	100.0	90.0	94.7
	ThinkCentre M51 Desktop	9	10	9	100.0	90.0	94.7
	Sum/average	94	103	94	100.0	91.3	95.5
(b) Unix Server							
HP	RX8620	18	18	18	100.0	100.0	100.0
	RX7620	16	16	16	100.0	100.0	100.0
	RX4640	25	25	25	100.0	100.0	100.0
	RP8420	19	16	16	84.2	100.0	91.4
	RP7420	18	15	15	83.3	100.0	90.9
IBM	P5 590	14	13	13	92.9	100.0	96.3
	P5 570	18	18	16	88.9	88.9	88.9
	P5 550	10	11	10	100.0	90.9	95.2
	P5 520	10	11	10	100.0	90.9	95.2
	P5 510	10	11	10	100.0	90.9	95.2
	Sum/average	158	154	149	95.0	96.2	95.6
(c) Monitor							
HP	vs1717 Flat-Panel LCD Monitor	10	11	10	100.0	90.9	95.2
	f1905 19 LCD Flat-Panel Monitor	11	11	11	100.0	100.0	100.0
	Pavilion mx704 17 Flat-Screen Monitor	12	11	11	91.7	100.0	95.7
	s7540 CRT Monitor	13	11	11	84.6	100.0	91.7
IBM	ThinkVision L170p Monitor	25	27	24	96.0	88.9	92.3
	ThinkVision L150 Monitor	20	19	19	95.0	100.0	97.4
	E74M 17 inch CRT Monitor	21	25	21	100.0	84.0	91.3
	E54 15" Monitor MPRII	11	10	10	90.9	100.0	95.2
	Sum/average	123	125	117	95.1	93.6	94.3
(d) Printer							
HP	Inkjet 1200dn printer	17	21	17	100.0	81.0	89.5
	DeskJet 3845 Printer	6	6	6	100.0	100.0	100.0
	Color LaserJet 2600n Printer	10	11	10	100.0	90.9	95.2
IBM	Infoprint Color 1357 Printer	6	5	5	83.3	100.0	90.9
	Infoprint 1412 Printer	10	9	9	90.0	100.0	94.7
	Infoprint 1352 Printer	9	8	8	88.9	100.0	94.1
	Sum/average	58	60	55	94.8	91.7	93.2
Total/summary		433	442	418	96.5	94.6	93.2

Table III.
The results of performance measurement for ITSIES

Notes: No. TIE: number of target information entities; No. IEE: number of information entities extracted; No. CIE: number of correct information entities extracted

Concerning the generality on the ITSIES system, this study collects twenty-two products from other companies' Web pages as a testing dataset to compare the extraction results of JAPE rules. Two criteria exist for adopting IT products from different computer companies – i.e. representation and heterogeneity. The former indicates the specific product that is well-known in a famous company, e.g. Sun's Unix server and Epson's printer. The latter one aims at observing the difference in specification representation for different countries, e.g. Sony and Canon in Japan, and Asustek, Acer, and BenQ in Taiwan. Table IV shows the evaluation results.

The F-measures are over 50 per cent for PC and Monitor, which indicate that the ITSIES system has a better performance for evaluating the two products. Although Sony, Asustek, Acer, and BenQ are all located in Japan or Taiwan, their product

Company	Product name	Total tokens	No. TIE	No. IEE	No. CIE	Recall (%)	Precision (%)	F-measure (%)
(a) Personal computer								
Sony	VGN-T370P-L Notebooks	3,775	9	10	7	78.0	70.0	73.8
	VGN-FS675P-H Notebook	3,649	9	8	6	67.0	75.0	70.8
	VAIO V167G TV-PC Desktop	3,534	14	12	10	71.0	83.3	76.7
Asustek	W5A Notebook	549	10	6	5	50.0	83.3	62.5
	W3V Notebook	663	9	6	5	56.0	83.3	67.0
	V6V Notebook	531	10	6	5	50.0	83.3	62.5
	Sum/average		61	48	38	62.0	79.2	69.6
(b) Unix Server								
SUN	Fire V890 Server	1,234	13	2	2	15.4	100.0	26.7
	Fire V40z Server	1,087	15	10	8	53.3	80.0	64.0
	Fire V240 Server	1,172	13	5	5	38.5	100.0	55.6
	Fire V1280	950	16	4	4	25.0	100.0	40.0
	Sum/average		57	21	19	33.1	95.0	49.0
(c) Monitor								
BenQ	USA - FP531 LCD monitor	480	5	5	3	60.0	60.0	60.0
	FP71V LCD monitor	521	6	5	3	50.0	60.0	54.5
	FP71E LCD monitor	419	5	5	3	60.0	60.0	60.0
	FP537s LCD Monitor	467	6	5	3	50.0	75.0	60.0
Acer	AF715 CRT monitor	217	6	4	3	50.0	75.0	60.0
	AC501 CRT monitor	190	7	4	3	43.0	75.0	54.7
	Sum/average		35	28	18	51.0	64.3	56.9
(d) Printer								
Epson	PictureMate	1,458	3	1	1	33.3	100.0	50.0
	Stylus Photo R300 Printer	1,721	3	2	2	66.7	100.0	80.0
	Stylus C66 Printer	1,369	4	3	3	75.0	100.0	85.7
Canon	PIXMA iP3000 printer	1,033	6	1	1	16.7	100.0	28.6
	Printers - i80 Printer	1,022	9	2	2	22.2	100.0	36.3
	PIXMA iP90 Printer	1,128	6	1	1	16.7	100.0	28.6
	Sum/average		31	10	10	32.3	100.0	48.8
Total/summary			184	107	85	44.6	84.6	58.4

Notes: No. TIE: number of target information entities; No. IEE: number of information entities extracted; No. CIE: number of correct information entities extracted

Table IV.
The extraction results for testing other IT products

information is published at a high correspondence with standardized specifications for international businesses. On the contrary, the F-measures for a Unix server and Printer are below 50 per cent. This is a warning sign for evaluating these products, as system tailoring is necessary in advance. Specifically, the recall in Canon's Printer shows that the worst value is below 25 per cent. A possible explanation might be on the difference of language representation and coding skill, or even culture. Advanced exploration could be necessary for future research.

To summarize up, the results of the average F-measure (58.4 per cent) indicate an acceptable performance. The average precision is above 84 per cent, but the average recall value goes down to 44 per cent. Obviously, the recall decreases sharply and the precision still remains at a reasonable level along with the extension of a Web site scope. This finding implies that the JAPE grammar rules in the ITSIES system are acceptable, but need to be improved.

Conclusion and contribution

Information extraction is an important task in knowledge management from the process perspective. In this paper we present an approach that combines natural language processing and the ontology concept to extract information for unstructured IT product information. It also builds up a system (ITSIES) with JAPE grammar rules to examine the performance.

The major challenge of an information extraction methodology is how to enhance system quality by performance measurement. From the extraction results, we see that the JAPE grammar rules have good performance for a specific knowledge domain in IT products. This approach not only proposes an information extraction architecture, but also links up with NLP technology and presents the extraction results with ontology language. However, product Web pages are not uniformly structured and IT specification information is coded in diverse forms, leaving many specific limitations that still need to be overcome. More empirical studies are needed in order to validate the effectiveness of the system, and the methods for extracting information can be improved continuously.

References

Alani, H., Kim, S., Millard, D.E., Weal, M.J., Hall, W., Lewis, P.H. and Shadbolt, N.R. (2003), "Automatic ontology-based knowledge extraction from the web documents", *IEEE Intelligent Systems*, Vol. 18 No. 1, pp. 14-21.

Bontcheva, K., Maynard, D., Tablan, V. and Cunningham, H. (2003), "Gate: a unicode-based infrastructure supporting multilingual information extraction", paper presented at IESL03: Workshop on Information Extraction for Slavonic and other Central and Eastern European Languages, held in conjunction with the 4th International Conference "Recent advances in natural language processing" (RANLP 2003), Borovets.

Chung, W., Chen, H. and Nunamaker, J.F. (2005), "A visual framework for knowledge discovery on the web: an empirical study of business intelligence exploration", *Journal of Management Information Systems*, Vol. 21 No. 4, pp. 57-84.

Cowie, J. and Lehnert, W. (1996), "Information extraction", *Communications of the ACM*, Vol. 39 No. 1, pp. 80-91.

De Sitter, A., Calders, T. and Daelemans, W. (2004), "A formal framework for evaluation of information extraction", Technical Report TR 2004-0, Department of Mathematics and Computer Science, University of Antwerp, available at: www.cnts.ua.ac.be/Publications/

Edgington, T., Choi, B., Henson, K., Raghu, T.S. and Vinze, A. (2004), "Adopting ontology to facilitate knowledge sharing", *Communications of the ACM*, Vol. 47 No. 11, pp. 85-90.

Embley, D.W., Tao, C. and Liddle, S.W. (2005), "Automating the extraction of data from HTML tables with unknown structure", *Data & Knowledge Engineering*, Vol. 54, pp. 3-28.

Embley, D.W., Campbell, D.M., Jiang, Y.S., Liddle, S.W., Lonsdale, D.W., Ng, Y.K. and Smith, R.D. (1999), "Conceptual-model-based data extraction from multiple record web pages", *Data & Knowledge Engineering*, Vol. 31, pp. 227-51.

Jung, H., Yi, E., Kim, D. and Lee, G.G. (2005), "Information extraction with automatic knowledge expansion", *Information Processing and Management*, Vol. 41, pp. 217-42.

Khelif, K. and Dieng-Kuntz, R. (2004), "Ontology-base semantic annotation for biochip domain", paper presented at KMOM Workshop ECAI 2004, Valencia.

Lavelli, A., Califf, M., Ciravegna, F., Freitag, D., Giuliano, C., Kushmerick, N. and Romano, L. (2004), "IE evaluation: criticisms and recommendations", paper presented at ATEM-2004: AAAI-04 Workshop on Adaptive Text Extraction and Mining, San Jose, CA.

Liebowitz, J. (2001), "Knowledge management and its link to artificial intelligence", *Expert Systems with Applications*, Vol. 20, pp. 1-6.

Liu, L., Pu, C. and Han, W. (2001), "An XML-enabled data extraction toolkit for web sources", *Information Systems*, Vol. 26, pp. 563-83.

Maynard, D. and Cunningham, H. (2003), "Multilingual adaptations of a reusable information extraction tool", *Proceedings of the Demo Sessions of EACL'03, Budapest*.

Navigli, R., Velardi, P., Cucchiarelli, A. and Neri, F. (2004), "Quantitative and qualitative evaluation of the OntoLearn ontology learning system", paper presented at ECAI Workshop on Ontology Learning and Population, Valencia.

Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D. and Goranov, M. (2003), "KIM – Semantic Annotation Platform", *Proceedings of the 2nd International Semantic Web Conference, Sanibel Island, FL*.

Soderland, S. (1999), "Learning information extraction rules for semi-structured and free text", *Machine Learning*, Vol. 34, pp. 233-72.

Strzalkowski, T. and Vauthey, B. (1992), "Information retrieval using robust natural language processing", *ACL'92 Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, University of Delaware, Newark, DE*, pp. 104-11.

Sung, N.H. and Chang, Y.S. (2004), "Business information extraction from semi-structured webpages", *Expert Systems with Applications*, Vol. 26, pp. 575-82.

Thirunarayan, K., Berkovich, A. and Sokol, D.Z. (2005), "An information extraction approach to reorganizing and summarizing specifications", *Information and Software Technology*, Vol. 47, pp. 215-32.

Thompson, C.A., Califf, M.E. and Mooney, R.J. (1999), "Active learning for natural language parsing and information extraction", *Proceedings of the 16th International Conference on Machine Learning, Bled, June*.

About the authors

Chyan Yang received his PhD in computer science from the University of Washington, Seattle, USA. He also holds a MS in Information and Computer Science from Georgia Institute of Technology, USA, an MBA in Management Science from National Chiao Tung University.

Professor received his B.S. in EE at National Chiao Tung University. Between 1987 and 1992 he worked as an assistant professor in the Department of Electrical and Computer Engineering, US Naval Postgraduate School at Monterey, California. From 1992 to 1995 he was with the Institute of Management Science, National Chiao Tung University, Taiwan as an associate professor. Chyan Yang is now a Professor in the Institute of Business and Management, and Director of EMBA program at National Chiao Tung University, Taiwan. Professor Yang has published more than 60 journal papers and 90 conference papers. His researches have been published in various SCI and SSCI journals including *The Electronic Library, Journal of the American Society for Information Science and Technology, International Journal on Information Management, International Journal of Computer Communications*, and *IEEE Journal on Selected Areas on Communication*. His current researches include information management, technology management, and strategic management. Professor Yang worked as an advisor to several IT companies. Chyan Yang is the corresponding author and can be contacted at: professor.yang@gmail.com

Liang-Chu Chen is an assistant professor at the Department of Information Management, Management College, National Defense University, Taiwan. He received his MS in the Institute of Resource Management from National Defense College, and a PhD in the Institute of Information Management, National Chiao-Tung University, Taiwan. His research interests include knowledge management, information extraction and data mining.

Chun-Yen Peng is a senior IT engineer of Taiwan Semiconductor Manufacturing Company (TSMC) and graduate from Institute of Information Management, National Chiao Tung University on 2005. From 1995 to 2001 he was MIS engineer of Acer Computer group. He supported Acer Computer group to build AGN (Acer Global Network) and mail system to handle message exchange among worlds wide over 200 branch offices. He has extensive experience on IT infrastructure system. His current job is IT infrastructure architecture design and project management for TSMC.