2010 TOPCO 崇越論文大賞

論文題目：

Designing a Cross-language Comparison Shopping Agent

報名編號： C0016

Abstract

This research pertains to the design and development of a shopbot called WebShopper+. This shopbot is intended to help customers find and compare e-tailers that market their wares using different languages. WebShopper+ is built with a multilingual ontology to overcome the language barriers that arise with global e-commerce. This research proposes a semiautomatic method of constructing a multilingual ontology by using the formal concept analysis and association analysis. It also proposes an automatic method for the categorization of product data into predefined classes, with the aim of alleviating administrators' task load. Additionally, a semantic search mechanism based on concept similarity is designed to assist customers in finding more desirable products. The experimental results show that these methods perform well and the shopbot can help customers find real bargains on the Web and to find products that cannot be bought locally.

Keywords: Shopbot, Comparison shopping, Ontology, Semantic similarity, Formal concept analysis.

1. Introduction

Online purchase-decision aids are becoming more and more important while the number of e-tailers is increasing. Consumers may purchase products through shopbots that can collect useful information from numerous e-tailers and present the data in a meaningful way, providing customers with purchase-decision aids. Shopbots help customers to compare these vendors so consumers can make better purchase decisions and subsequently purchase from or place a bid with suitable vendors. In addition, they enable vendors to monitor their competition and to reach a larger customer base (Fasli, 2006; Huang and Tsai, 2009). Beyond these fundamental services, some researchers have attempted to expand the service horizon of shopbots by focusing on the utility of consumer purchasing behavior (Montgomery, Hosanagar, Krishnan, and Clay, 2004), identification of the best price for a bundle of items (Garfinkel, Gopal, Tripathi, and Yin, 2006), and integration of sales promotion information into search results (Garfinkel, Gopal, Pathak, and Yin, 2008). Because most electronic purchases are still not automated, customers often expend excessive amounts of time in the buying process, which includes collection and interpretation of information on many vendors and products, making purchase decisions, and subsequently entering purchase and payment information. Software agent technologies such as a shopbot provide a basis to solve these issues as they automate the most time-consuming activities of the purchasing process (Maes, Guttman, and Moukas, 1999).

Nevertheless, existing shopbots are only designed to collect information from e-tailers that use common language, or e-tailers that reside in a single nation. As a result, many customers are unable to locate real bargains on the Web and many e-tailers are constrained from reaching much of their potential market (Huang and Tsai, 2009). Previous researches have showed that the trade between countries that share a common language is three times greater than trade between countries without a common language (Ghemawat, 2001). In addition, Internet users tend to surf websites presented in their native language (Grace-Farfaglia et al., 2006; Lynch and Beck, 2001). To solve this problem, Huang and Tsai developed a shopbot with a multilingual ontology that would allow customers use their native language in searching for online product catalogs that were presented in different languages. Their experimental result showed that customers can locate more products and find a greater number of bargains over the Web, using their shopbot (Huang and Tsai, 2009). A multilingual ontology is the taxonomy (i.e., a hierarchical structure of classifications for a given set of objects) of products, in which categorizations can be expressed in different languages and their subsumption and equivalence relationships can be defined. However, system administrators have to build this ontology manually and define product-classification rules before the shopbot can collect product data from venders' websites. Since product information provided by vendors is massive and changes frequently and defining relationships among thousands of classes in different languages would be very time-consuming, more automatic approaches are required for building ontology and classifying product data efficiently. Moreover, Huang and Tsai have developed a semantic searching mechanism that can interpret equivalence and subsumption relationships between concepts described in different languages (Huang and Tsai, 2009). However, this searching mechanism only considers equivalence and subsumption; it does not consider concept similarity.

This research aims to design and develop a shopbot that can help customers to compare products located in e-stores, using different languages. To overcome the shortcomings of existing shopbots, this research (1) proposes a semi-automatic method of constructing a multilingual ontology; (2) designs an automatic method of classifying product data into said ontology; and (3) proposes a semantic searching mechanism based on concept similarity.

2. Related Work

This section introduces the conceptualization of shopbot and comparison shopping site. Additionally, the technical foundations adopted to develop our shopbot are discussed.

2.1. Shopbots and Comparison-Shopping Sites

Shopbots, also known as comparison-shopping agents, are automated tools that query e-commerce sites, such as online shops, to retrieve product information. They then parse the received information to extract useful product and vendor information, which can be used to aid customers in making purchase decisions. These agents employ an automatic process to build wrappers and they utilize a number of rules to parse semi-structured Web pages (Doorenbos, Etzioni, and Weld, 1997). Shopbots are one type of specialized agents that are designed to help users filter and process information, by retrieving product details and comments, comparing products, vendors, and services based on user-defined criteria, searching for products or services of the best-value, monitoring product availability or special offers and discounts from online shops, recommending services and products, and identifying new products of potential interest (Fasli, 2006).

In contrast to shopbots, comparison-shopping sites do not rely on agent-based technology. These sites depend on vendors to provide a required set of information, or they operate as meta-search engines of vendor sites (Fasli, 2006). Customers can obtain a search result in the form of a list of items, with prices from different shops, and they can then identify the shop that offers the best price. The customers can then click the link they choose, in order to purchase from the associated shop.

Reviewing the most popular comparison-shopping sites (eBizMBA, 2009), it is readily apparent that existing sites only support keyword search and do not support semantic search. A keyword search is unable to process relationships between keywords and therefore cannot support searches in multiple languages, even if the keywords have the same meaning in different languages. Moreover, existing shopbots and comparison-shopping sites only collect information from e-retailers that use a common language, or that reside in the same nation, thus customers are often not able to locate many of the bargains that exist on the Web and e-retailers are incapable of reaching much of their potential customer base around the world.

2.2. Formal Concept Analysis

This study presents the design of a method for semi-automatic ontology construction and a corresponding approach to automatic data classification, based on the formal concept analysis (FCA). The methodology of the FCA was first proposed by Rudolf Wille (Wille, 1982). It is often applied to the construction of ontologies that are based on large data sets. FCA usually relies on concept lattices to construct domain ontologies (Formica, 2006; Haav, 2004; Stumme and Maedche, 2001; Tho, Hu, Fong, and Cao, 2006). Ontology, in the philosophical view, refers to a discipline that deals with the nature and the organization of being. In this sense, we can refer to an ontology as a particular system of categorizations that reflects a given perspective of the world. In computer science, an ontology refers to an engineering artifact consisting of (1) a specific vocabulary, which appears as concepts and relations that are used to describe a certain reality and (2) a set of explicit assumptions regarding the intended meaning of the vocabulary (Maedche, 2002).

In general, the process of FCA-based ontology construction entails extracting keywords from data sources, constructing a formal context, and constructing concept lattices (Chen and Wu, 2008; Weng, Tsai, Liu, and Hsu, 2006). First, during the keyword extraction phase, a number of keywords are extracted from documents. Second, during the formal context construction phase, the relationships between keywords and documents are established using a table. Table 1 provides an example of a formal context, in which “X” denotes that the keyword occurs in the corresponding document. Finally, in the concept lattice building phase, the formal context is transformed into a concept hierarchy based on two relationships: inheritance and intersection. These relationships are defined as follows.

1. Inheritance: when documents that contain keyword *A* also contain keyword *B*, but not all documents that contain keyword *B* necessarily contain keyword *A*, concept *A* is said to be a sub-concept of *B*, expressed as $A \subseteq B$.
2. Intersection: when some documents that contain keyword *A* also contain keyword *B*, and vice versa, the concepts *A* and *B* are said to have an intersection relationship.

Table 1. Example of formal context

	E-Commerce	**B2B**	**B2C**	**C2C**	**SCM**	**CRM**
Document 1	X	X	X	X		
Document 2	X	X			X	
Document 3	X	X	X			
Document 4	X		X	X		X
Document 5	X			X		
Document 6	X	X	X	X		
Document 7	X		X			X

Thus, we can transform the formal context expressed in Table 1 into a visualized relationship diagram, referred to as a concept lattice, as in Fig. 1. In this figure, arrows denote an inheritance relationship and dashed lines represent an intersection relationship. This concept lattice represents an example of an ontology.

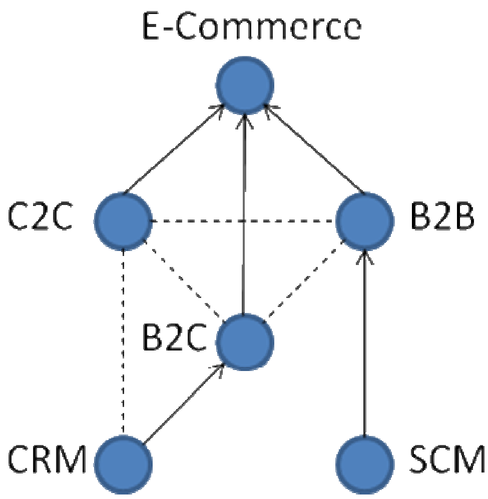

Fig. 1. Example of concept lattice

2.3. Semantic Similarity

Lin defines a number of intuitive concepts pertaining to similarity, described as follows (Lin, 1998):

1. Intuition 1: The *commonality* between concepts *A* and *B* determines their similarity. The greater the commonality they share, the more similar they are.
2. Intuition 2: The *differences* between *A* and *B* determine their similarity. That is, the less the differences between them, the more similar they are.
3. Intuition 3: When both *A* and *B* are *identical*, the maximum degree of similarity between *A* and *B* is reached, regardless of the commonality they share.

Semantic similarity also refers to similarity between two concepts in an “is-a” taxonomy, such as WordNet, and reflects the extent to which they share information in common (Resnik, 1995). For instance, the concept of a “personal computer,” in English, and the concept of “個人電腦,” in Chinese, have the same meaning, but they are depicted using different syntax. Furthermore, we can say that “pizza” and “food” are more closely related than “pizza” and “drink” because the former are more similar in meanings than the latter.

Resnik (1995) proposes an approach to measuring semantic similarity based on information theory. The more information two concepts share in common, the more similar they are considered to be. The information shared by two concepts also reflects the information content of those concepts that subsume them in the taxonomy, that is,

$$sim_{\text{Resnik}}(C_1, C_2) = max\, I(common(C_1, C_2)) = -\log P(C_0) \quad , \quad (5)$$

where C_o is the most specific common super-class between C_1 and C_2, and $P(C_o)$ is the probability that a randomly selected object belongs to C_o.

Lin (1998) similarly proposed a method that presents an information theoretic definition of similarity that can be used in every domain with a probabilistic model. This method is suitable for measurement of the semantic similarity between two concepts in an "is-a" taxonomy. Lin used the notation $sim(C_1, C_2)$ to denote the similarity between x_1 and x_2, where $x_1 \in C_1$, $x_2 \in C_2$, and the selection of a generic C_1 is not related to the selection of a generic C_2. The amount of information included in "$x_1 \in C_1$ and $x_2 \in C_2$" is

$$-\log P(C_1) - \log P(C_2), \tag{2}$$

where $P(C_1)$ and $P(C_2)$ are probabilities that a randomly selected instance belongs to the concepts C_1 and C_2, respectively. Supposing the taxonomy is a tree-like hierarchy, and C_0 is the most specific super-concept of C_1 and C_2, the commonality between x_1 and x_2 is $x_1 \in C_0$ and $x_2 \in C_0$. As such, the similarity is expressed by

$$sim(x_1, x_2) = 2 \times \frac{\log P(C_0)}{\log P(C_1) + \log P(C_2)}. \tag{3}$$

The ontology-distance approach may cause problems, however, if there are two concepts having a great distance between them due to the ontology hierarchy. When using the distance approach, such concepts may be determined to be less similar than is actually the case. The ontology-distance approach tends to produce a similarity calculation that is biased downward. The results of an experimental evaluation have shown that Lin's method produces output that is closer to true human judgments than does Rensik's method or the ontology-distance approach (Lin, 1998). As such, this research adopts Lin's method to calculate semantic similarity and applies it to the design of a semantic searching mechanism.

3. System Architecture

Fig. 2 illustrates the system architecture of a cross-language shopbot. A prototype system, termed WebShopper+, was developed based on this architecture. WebShopper+ helps customers to compare vendors and purchase costs of computer and business books.

The data collection agent collects product data from two types of Web documents: XML and HTML documents. XML documents are acquired from e-stores that offer Web services, while HTML documents are acquired from e-stores that do not offer Web services. After the data collection agent has collected and parsed all of the useful product data, it then classifies the data according to a predefined product ontology, calculates purchase costs according to list prices, delivery fees, and exchange rates, and then saves product related data into the product database. The purchase cost is calculated in a specific currency by the following equation:

$$Purchase\ cost = (list\ price + delivery\ fee) \times exchange\ rate. \quad (4)$$

The exchange rate agent updates exchange rates daily. Calculating purchase cost in a specific currency makes it easy for users to determine the prices they must pay to purchase products from different vendors and to find the most economical distributor.

Customers can employ the semantic searching mechanism to search for products in different languages. This prototype supports Chinese, English, and Japanese. The minimum similarity threshold can be set by the user to filter out those concepts that are less likely to be of interest. After searching, the mechanism will list and sort similar concepts by descending similarity. Users can then click one of the listed concepts to view the products associated with it. They can then compare the items further, in term of purchase cost, to find the best vendor.

System administrators use three different editors to maintain the product ontology, delivery fees, and exchange rates. Additionally, an ontology construction toolkit provides some tools to help administrators build the ontology in a semi-automatic manner.

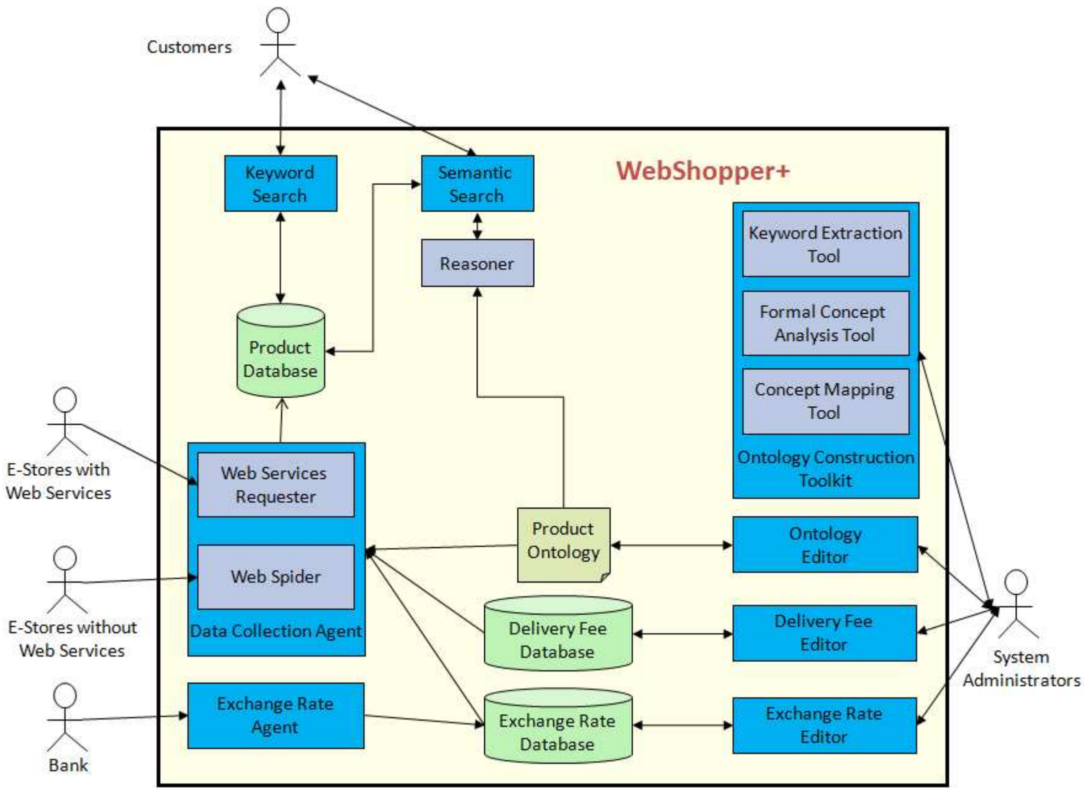

Fig. 2. System architecture of WebShopper+

4. Ontology Construction Method

This research proposes a novel approach to constructing a multilingual ontology and develops an ontology construction toolkit that includes a keyword extracting tool, FCA tool, and concept mapping tool, to help administrators build their product ontology in a semi-automatic manner. The following subsections will introduce the procedure and set of tools.

4.1. Ontology Construction Procedure

To efficiently construct a multilingual ontology that is capable of supporting product data classification and semantic search, a semi-automatic ontology construction procedure is proposed. The following steps detail this procedure:

1. Construct a classification tree for each vendor according to the taxonomy used by said vendor. This step assists administrators in building initial classification trees quickly.
2. Use the keyword extraction tools to obtain keywords from book titles, then remove meaningless or infrequent words to form the keyword set.
3. Use the FCA tool to analyze the context of the book titles and keywords generated in Step 2, and further extend the classification trees produced in Step 1. For example, suppose there is a conceptual term, "Java," in an original classification tree and that the FCA tool has found that an inheritance relationship exists between this concept and "JavaServer Pages." If this inheritance relationship is not revealed in the original classification tree, the administrator is advised to extend the classification tree by adding the concept "JavaServer Pages" as a sub-concept of "Java."
4. Combine all extended classification trees that use common languages into a single integrated one; this step generates a classification tree for each language.
5. Combine the trees, in different languages, into the final ontology. First, the concept-mapping tool is used to translate non-English concepts into English. Second, compare these translated concepts with the concepts in the English tree. If there are two concepts in different trees that are equivalent, these two concepts, in their original languages, are defined to have an equivalence relationship. If no equivalence relationship is defined, their sub-concept words are then compared. If the majority of sub-concepts in the different trees are equivalent, these two concepts may have an equivalence relationship, thus they require manual review by the administrator. These trees, in different languages, are ultimately combined into an ontology according to the equivalence relationships between their concepts.

By following the above steps, a multilingual ontology can be generated to assist the shopbot in (1) classifying book data from bookstores that use different languages and (2) providing a cross-language semantic searching mechanism.

4.2. Keyword Extraction Tool

The keyword extraction tool uses an *n*-gram method as well as part-of-speech taggers to extract keywords from product titles. An *n*-gram is a sub-sequence of *n* items taken from a larger sequence (Cohen, 1997). It is typically used in various fields of statistical natural language processing. The keyword extraction tool uses the word-level *n*-gram method to extract keywords from product titles. For example, the book title “Learning Java programming language” can be divided into the keywords “Java” and “language” using a 1-gram method, “programming language” using a 2-gram method, and “Java programming language” using a 3-gram method. There could be a large volume of keywords produced, thus the tool will attempt to ignore frequent nouns and to remove meaningless words.

The tool uses the CKIP part-of-speech tagger to tag the part-of-speech for words in Chinese book titles. In addition, it uses Mecab, which is a part-of-speech and morphological analyzer to deal with Japanese book titles, and it uses OpenNLP to tag the part-of-speech for words in English book titles. After tagging, the tool extracts nouns, using 1, 2 or 3-grams to generate the candidate conceptual words that will be used in the subsequent FCA.

4.3. Formal Concept Analysis Tool

To perform FCA, the FCA tool generates the formal context representing the relationship between product titles (objects) and their keywords (attributes). It then applies association analysis to determine the inheritance relationships between concepts. Association analysis is a technique developed in the field of data mining; it can find co-occurrence patterns that are less obvious, from a large data source. In identifying association rules, association analysis depends on the measures of *support* and *confidence*. The FCA tool employs association analysis to find inheritance relationships between concepts. First, the association analysis generates a frequent 2-item set that reflects keyword pairs that frequently co-occur in product titles. For example, if the keyword “programming language” usually appears with the keyword “Java” we can likely conclude that the concepts “programming language” and “Java” have an intersection or inheritance relationship. Second, the confidence of the association rules “Java → programming language” and “programming language → Java” are calculated. If a confidence of an association rule, such as “Java → programming language”, is equal to or slightly less than 1 we can likely conclude that programming language is a super-concept of Java, according to the theory of FCA. RapidMiner is used to perform the association analysis in the FCA process. We define the minimum confidence to be 0.9 and the minimum support to be 0.6 in generating the association rules. Classification trees are extended according to these rules.

4.4. Concept-Mapping Tool

To combine classification trees from different languages into a single ontology, we use English as the pivot language for translations between different languages. The concept-mapping tool uses the Google Dictionary API and the Wikipedia API to perform the automatic translation. First, the tool uses the Google Dictionary API to access the Japanese-English and Chinese-English dictionaries, to automatically translate concepts presented in Japanese or Chinese into English. However, using bilingual dictionaries does not allow us to address some words, such as proper nouns and abbreviations. To tackle this problem, the tool also accesses the Wikipedia API to deal with conceptual words that cannot be translated into English simply by looking to a bilingual dictionary. Those words that cannot be automatically translated using the Google Dictionary or Wikipedia APIs are then manually removed, revised or translated by an administrator. This is done, for example, by using Yahoo! Kimo Dictionary, Sanseido Web Dictionary or Yahoo! Japan Dictionary. After completion of the above steps, the tool detects which concepts are equivalent between different languages and helps the administrator to combine the classification trees from each language into the final ontology.

5. Data Collection Agent and Data Classification Method

The data collection agent retrieves product data from e-stores with Web service or HTML format. The data collection agent uses a Web service request to query the Web services provided by e-stores. First, the agent requests product data via a REST (Representation State Transfer) style Web service. Then, the agent receives a number of XML documents in reply, which include the product data, and subsequently parses them to extract useful information.

However, there are still many e-stores that do not provide these Web services. The data collection agent uses a tool referred to as a “Web spider” to retrieve Web pages in HTML format. The freeware tool HTTrack is an example of a typical Web spider. It crawls across Web pages based on some initial seeding data (starting pages), and retrieves the associated pages that match some predefined conditions.

After product pages are collected, the data collection agent parses these pages and extracts useful product information. The agent automatically classifies this product data as per the predefined categorizations in the product ontology. The classification procedure is described as follows:

1. The agent collects product data from online bookstores and identifies book titles, as well as the books’ original categories, as defined by the bookstores.
2. The agent then searches within the ontology for the concepts that match with the original category of the product. Because the concepts in the product ontology are sourced from keywords in book titles and the original book categorizations, all book data can be mapped to a corresponding concept.
3. Finally, the agent checks if there are sub-concepts that are subsumed by the

corresponding concept in the ontology. It then finds the most specific concept that is contained in the book title and uses this to determine the classification of the book data.

When the agent collects product data, it will also update the existing data in its product database in a period of time.

6. Semantic Search

Semantic search is designed to help customers find products that have a high semantic similarity with some specified concept. For example, the concept of “conveyance” could also refer to “transportation,” or “vehicle,” and may include the sub-concept of “train.” If we type “conveyance” into a keyword search, we will only retrieve those items that include the keyword “conveyance” in their title. A keyword search mechanism cannot identify related results like “transportation,” “vehicle,” or “train,” even though they share similar meaning. Further, we are unable to obtain any results pertaining to the concept of “conveyance” by typing the keyword “交通工具,” in Chinese, even though the meaning is identical. To address these gaps, this system provides a semantic search engine that will assist users in finding products based on the semantic meaning of specified concepts.

When a customer inputs a conceptual word and specifies the minimum semantic similarity to consider, the data is transmitted to the searching mechanism, which is developed using the Jena API. The mechanism will reason, based on the ontology, and attempt to determine if there is a consistent concept amongst the search terms. The system will then begin to calculate the semantic similarities between the specified concepts and other concepts in the ontology, using Lin’s method (Lin, 1998). Finally, the concepts that have a similarity greater than the user-defined parameter are returned. This semantic searching mechanism differs from the mechanism proposed by previous research (Huang and Tsai, 2009). Previous work has only considered sub-concepts of a queried concept. But in this research, the semantic searching mechanism considers all concepts having high similarity with the queried concept. In other words, the search range can be viewed as a circle, where the center of a circle is the queried concept and the radius is determined by the user-defined minimum semantic similarity.

7. System Evaluation

This research proposes a prototype system, referred to as WebShopper+, which is developed based on the proposed architecture. WebShopper+ is intended to help customers search for computer books and business books, and to aid the comparison of their purchase costs and their respective vendors. WebShopper+ collects book data from the most popular online book-sellers, including Amazon (America, England, and Japan), Yahoo Shopping (Japan), Books (Taiwan), and Eslite (Taiwan) (note that Books and

Eslite do not provide Web services). This shopbot deals with new books only and excludes used books. The bot is implemented using the Java programming language, with the Jena API.

Users are able to use their native languages to search for books. For example, a Taiwanese customer can enter the conceptual word “3D 程式設計,” which means 3D programming, and can choose 0.7 as the minimum semantic similarity for the search results. Fig. 3 depicts the search results for this query, where the listed concepts are similar to “3D 程式設計,” with the similarity score provided (in parentheses). The top 10 concepts returned are 3D 程式設計 (3D programming), 應用程式 (applications), 繪圖與多媒體 (graphics and multimedia), Adobe, 簡報 (presentation), 3D, Adobe Acrobat, 數位攝影 (digital photography), 圖像處理 & 創作 (image manipulation & creation), and 3D 繪圖 (3D graphics).

Fig. 3. System interface

The customer clicks the concept they find most interesting and all books belonging to this category are returned (see Fig. 4). As the user locates the book that he needs, he can press the button “比價” (compare prices) to check which vendors sell this product and at what prices, in order to find the best vendor. Fig. 5 shows, in this case, that the first item sold by Amazon in the United Kingdom is the cheapest book (NTD 1171) and the last item sold by Books in Taiwan is the most expensive one (NTD 1813). In this case, WebShopper+ enables the customer to use Chinese to search for English books and to reach foreign vendors. Thus, the customer saves NTD 642 if he purchases the book from Amazon in the UK, compared to Books, which is located in Taiwan. The customer can visit the vendor’s page by clicking the book title and can then purchase the book.


書籍查詢結果:3D 程式設計

圖示	書名	作者	ISBN13	ISBN10	比價
	3D Game Engine Design : A Practical Approach to Real-Time Computer Graphics (The Morgan Kaufmann Series in Interactive 3D Technology)	David H. Eberly	9781558605930	1558605932	比價
	3D Game Engine Design: A Practical Approach to Real-Time Computer Graphics (The Morgan Kaufmann Series in Interactive 3D Technology)	David H. Eberly	9780122290633	0122290631	比價
	3D Game Engine Programming (Game Development Series)	Oliver Duvel	9781592003518	1592003516	比價
	3D Game Environments: Create Professional 3D Game Worlds	Luke Ahearn	9780240808956	0240808959	比價
	3D Game Programming All in One (Course Technology PTR Game Development Series)	Kenneth Finney	9781592001361	159200136X	比價

Fig. 4. Books belonging to specified concept


此書籍有下列賣家:

圖示	書名	作者	原有價格	換算後價格 (原價+運費)*匯率	賣家
	3D Game Environments: Create Professional 3D Game Worlds	Luke Ahearn	GBP27.54	NTD1171.272	Amazon (UK)
	3D Game Environments: Create Professional 3D Game Worlds	Luke Ahearn	GBP27.54	NTD1237.2219	Amazon (US)
	3D Game Environments: Create Professional 3D Game Worlds	Luke Ahearn	USD37.96	NTD1580.8214	Amazon (US)
	3D Game Environments: Create Professional 3D Game Worlds	Luke Ahearn	NT1748.0	NTD1813.0	Books (TW)

Fig. 5. Comparison of vendors of similar products

7.1. Evaluation of the Ontology Construction and Data Classification Methods

We now present a performance evaluation of the ontology construction method, measuring its precision and coverage. The system collected book data from e-tailers prior to April 1, 2009. In total, 1,213,628 pieces of data pertaining to 495,787 distinct books was collected. The ontology (accessible at http://sites.google.com/site/shiulihuang/files/WebshopperPlus.rar) was constructed based on this data set, using the proposed ontology construction method and tools.

Four domain experts were invited to participate in the evaluation. The ontology was divided into two sections: computer books and business books. Two of the experts had pursued graduate studies in Information Management and Computer Science, respectively, and were therefore tasked with evaluating the ontology of computer books. The other two experts had pursued graduate studies in Information Management and Marketing Management, respectively, thus they were responsible for evaluating the ontology of business books.

After the domain experts had determined the number of misplaced and misnamed concepts, the precision of the ontology construction method was calculated by the following equation:

$$Precision\ of\ ontology\ construction = \frac{\text{The number of accurate concepts}}{\text{The number of concepts in ontology}}. \quad (5)$$

The experts also determined the coverage of the semi-automatic ontology construction method. The coverage represents the completeness of the ontology and can be calculated by the following equation:

$$Coverage\ of\ ontology\ construction = \frac{\text{The number of concepts in ontology}}{\text{The number of concepts that should be included in ontology}}. \quad (6)$$

Thus, we can evaluate whether the ontology, built using our semi-automatic ontology construction method, is comparable to the ontology that was manually constructed by experts.

The ontology pertaining to computer books contained 1,792 concepts. The degrees of precision of the ontology, as reported by the two relevant domain experts, were 99.944% and 95.647%, with an average of 97.796%. The coverage values reported were 99.666% and 99.335%, with an average of 99.501%. The ontology pertaining to business books contained 859 concepts. The precision values reported by the two relevant experts were 98.254% and 97.090%, with an average of 97.672%. The coverage values reported were 100% and 99.768%, with an average of 99.884%. These results suggest that the proposed ontology construction method is capable of achieving a degree of precision and coverage. Our semi-automatic method performs very well (comparable to experts in the field) at a very low cost. In addition, our method requires less time and a lower human task load than the manual construction of the ontology.

In order to assess the performance of the automatic classification method, we measure the precision of the classification method. The system re-collected book data from the same online book stores during the period from April 1 to May 15, 2009. In addition to the original book data collected prior to April 1, 3,102 new pieces of data were collected on 723 distinct books. We randomly selected 100 books from the product database, and invited a domain expert (i.e., a graduate student that had majored in Information Management) to determine whether this new data was classified into appropriate classes. The precision of the data classification method is calculated as follows:

$$Precision\ of\ classification\ method = \frac{\text{The number of books that are classified accurately}}{\text{The number of randomly selected books}}. \quad (7)$$

The precision of the classification method was determined to be 100%. This result indicates that this method is capable of automatically, and correctly, classifying book data into the concepts predefined in the ontology.

7.2. Evaluation of the Semantic Searching Mechanism

In order to prove that the semantic searching mechanism is useful, able to satisfy users' demands, and is more effective than a search method that only considers sub-concepts, we conducted an Internet experiment. We invited Internet users to use a prototype of the proposed system, which incorporated two semantic searching mechanisms: searching for similar concepts and searching for sub-concepts. We asked users to search for any computer or business books, using each of the two mechanisms, and employing Chinese, Japanese, or English words as search terms. The subjects were then asked to fill out a questionnaire to measure their information and system satisfaction. We used metrics of information quality and system quality defined by DeLone and McLean's Information Systems Success Model (DeLone and McLean, 2003). The subjects' experiences with online shopping and background data were also collected as part of the questionnaire.

The invitation message was posted on the e-shopping and book forums of a bulletin board system (BBS) entitled PTT (ptt.cc), from June 30 to July 4, 2009. PTT is the largest BBS in Taiwan. There were 32 Internet users that participated in the experiment during this period. Table 2 summarizes the respondents' profiles. 37.5% of the respondents had experience purchasing items from foreign e-stores. The reasons for these purchases were that the products could only be bought from those other countries or that the products sold by foreign e-stores were cheaper. 50% of the subjects felt that the biggest difficulty in shopping at foreign e-stores was that they could not search for products using their native language. 46.9% of subjects indicated that their biggest difficulty in shopping at foreign sites was that they could not understand the foreign languages used. Therefore, language issues remain the biggest barriers to transacting at foreign e-stores.

Table 2. Demographic data of respondents

Measure	Items	Frequency	Percent
Experience in using a shopping comparison website	Yes	12	37.5%
	No	20	62.5%
Experience in purchasing from a foreign e-store	Yes	12	37.5%
	No	20	62.5%

Questions 1, 2, and 3 of the questionnaire were used to measure the information quality with respect to accuracy, completeness, and relevance. Questions 4, 5, and 6 were used to measure the system quality, in terms of functionality and importance. For questions 1 through 7, we used a 7-point Likert scale to measure the respondent's degree of agreement, where 1 represented extreme disagreement and 7 represented extreme agreement. We used a paired sample *t*-test to examine the difference in quality between the two systems, the results of which are shown in Table 3.

Table 3. Paired-samples *t*-test of systems with different searching mechanisms

Options	System 1 [Mean (SD)]	System 2 [Mean (SD)]	*t*-value (*p*-value)
1. Results of this searching mechanism are accurate	4.50 (1.295)	5.34 (1.285)	4.834*** (0.000)
2. Results of this searching mechanism are complete	4.66 (1.359)	5.09 (1.228)	2.239* (0.032)
3. Results of this searching mechanism are relevant	4.50 (1.459)	6.13 (0.609)	8.143*** (0.000)
4. This searching mechanism can assist me in finding the best offer	4.88 (1.008)	5.03 (1.062)	1.153 (0.258)
5. This searching mechanism can assist me in finding foreign sources of products	4.97 (1.177)	5.19 (0.931)	1.422 (0.165)
6. This searching mechanism is important for international comparison shopping	5.00 (1.136)	5.66 (0.937)	4.715*** (0.000)

*$p < 0.05$, ***$p < 0.001$; System 1: System considering sub-concepts, System 2: System considering similar concepts.

These result indicate that System 2, which incorporates the semantic searching mechanism addressing concept similarity, has better information quality in terms of accuracy, completeness, and relevance, versus System 1, which has a semantic searching mechanism that takes sub-concepts into consideration. With regard to system quality, the subjects indicated that System 2 was more useful for international comparison-shopping than System 1. However, the functionality of the two systems was not significantly different. The reason for this similarity may be that these systems used the same data sources and user interface. Overall, this result shows that the searching mechanism that considers conceptual similarities is more effective for international comparison shopping.

7.3. Evaluation of Shopbot

In order to show that the shopbot is useful, we assessed a random sample of 100 books to determine whether the majority of products tended to be distributed by e-stores sharing a common language, as well as whether differences in purchase costs existed among e-stores.

Table 4 shows the exchange rates and shipping fees used in this evaluation. The results show that two or more vendors sell the 56 books. The maximum difference in purchase cost was NTD 2750.495 (the book *Advances in Knowledge Discovery and Data Mining: 12th Pacific-Asia Conference,* the most expensive offer, was sold for NTD 7,220 by Books in Taiwan, while the cheapest offer was NTD 4,469.505, by Amazon in the United Kingdom) and the average difference between the highest and lowest purchase cost was NTD 399.806. This means that customers can save costs by using the shopbot. There are 29 books sold in two or more nations. Further, we found that 8 books were only sold in Taiwanese e-stores, 28 books were only sold in English e-stores, and 54 books were only sold in Japanese e-stores. These results reveal that cost variances do exist on the Web and that some products only exist in e-stores employing a certain language. The shopbot is able to help customers search for real bargains on the Web and to buy products that cannot be bought in their local countries.

Table 4. Exchange rates and shipping fees used in evaluation

Exchange rate	Shipping fee
USD 1 = NTD 32.975	Amazon US: USD 9.98
GBP 1 = NTD 53.83	Amazon UK: GBP 7.98
JPY 1 = NTD 0.3443	Amazon JP: JPY 2100
	Yahoo! Japan Shopping: JPY 1700

	Eslite: NTD 50
	Books: NTD 65

8. Conclusion

This research has proposed a design for a shopbot, termed WebShopper+, proposing a semi-automatic ontology construction method, an automatic data classification method, and a semantic searching mechanism for inclusion. This shopbot possesses a multilingual ontology, which can assist users in searching for products using their native languages. In addition, the proposed design also enables vendors to monitor their competitors and to reach more customers globally. The evaluation results suggest that the proposed ontology construction method is able to achieve high precision and coverage. The classification method can automatically classify product data correctly. Moreover, the shopbot is able to benefit customers by comparing purchase costs and product vendors that are located across different e-stores and that are using different languages. Using the shopbot, customers can locate real bargains on the Web and potentially purchase products that are not available in their own countries.

Since the language barrier, which poses an impediment to global e-commerce, still exists, the requirement for a cross-language comparison-shopping agent is becoming more apparent. A multilingual ontology will enable a shopbot to understand concepts in different languages, which not only addresses the language issue but also enables searching mechanisms in finding more suitable, relevant products. Although this prototype only addresses books and only supports Chinese, English, and Japanese languages, the system architecture can easily be expanded to support all human languages and all types of products. Moreover, the proposed ontology construction and classification methods save significant amount of time and resources while maintaining high accuracy. These methods can be employed in any context that entails ontology construction and data classification, such as document management systems and search engines.

The purchase decision-making process includes the following steps: need identification, information search, negotiation, purchase and delivery, and after-purchase service and evaluation (O'Keefe and McEachern, 1998; Turban et al., 2008). WebShopper+ supports the information search phase by answering the questions “what to buy?” and “from whom?” irrespective of the location of vendors or their languages. However, in order to overcome the language barrier, the other phases must also be supported. In the need identification phase, a recommender agent is required to proactively provide product information to customers on the basis of their profiles, preferences, and contexts. In the last three phases, related intelligent agents and intermediaries that can act as a proxy on the customer’s behalf to communicate with foreign vendors, to deal with international payments, deliveries, duties, and laws are also required. Thus, research that attempts to determine the best approach to the design of the agents and business models for these intermediaries is likely to prove fruitful.

References

Chen, D.-N. and Wu, C.-H., An Ontology-based Document Recommendation System: Design, Implementation, and Evaluation, *Proceedings of the Pacific Asia Conference on Information Systems*, Suzhou, China, 2008, pp. 469-480.

DeLone, W.H. and McLean, E.R., The DeLone and McLean Model of Information Systems Success: A 10-year Update, *Journal of Management Information Systems* 19 (4) (2003) 9-30.

Doorenbos, R.B., Etzioni, O., and Weld, D.S., A Scalable Comparison-shopping Agent for the World Wide Web, *Proceedings of the 1st International Conference on Autonomous Agents*, Marina del Rey, California, United States, 1997, pp. 39-48.

eBizMBA, Top 20 Comparison Shopping Websites, Retrieved August 1, 2009, from http://www.ebizmba.com/articles/shopping

Fasli, M., Shopbots: A Syntactic Present, a Semantic Future, *IEEE Internet Computing* 10 (6) (2006) 69-75.

Formica, A., Ontology-based Concept Similarity in Formal Concept Analysis, *Information Sciences* 176 (18) (2006) 2624-2641.

Garfinkel, R., Gopal, R., Tripathi, A., and Yin, F., Design of a Bundle Shopbot, *Decision Support Systems* 42 (3) (2006) 1974-1986.

Garfinkel, R., Gopal, R., Pathak, B., and Yin, F., Shopbot 2.0: Integrating Recommendations and Promotions with Comparison Shopping, *Decision Support Systems* 46 (1) (2008) 61-69.

Ghemawat, P., Distance Still Matters: The Hard Reality of Global Expansion, *Harvard Business Review* (September, 2001) 137-147.

Grace-Farfaglia, P., Dekkers, A., Sundararajan, B., Peters, L., and Park, S.-H., Multinational Web Uses and Gratifications: Measuring the Social Impact of Online Community Participation across National Boundaries, *Electronic Commerce Research* 6 (1) (2006) 75-101.

Haav, M.H., A Semi-automatic Method to Ontology Design by Using FCA, *Proceedings of 2nd Concept Lattices and their Applications*, Ostrava, Czech Republic, 2004, pp.

13-24.

Huang, S.-L. and Tsai, Y.-H., Developing a Shopbot with Multilingual Ontology for Global E-Commerce, *Journal of Internet Technology* 10 (2) (2009) 111-118.

Lin, D., An Information-theoretic Definition of Similarity, *Proceedings of the 15th International Conference on Machine Learning*, Madison, Wisconsin, USA, 1998, pp. 296-304.

Lynch, P.D. and Beck, J.C., Profiles of Internet Buyers in 20 Countries: Evidence for Region-specific Strategies, *Journal of International Business Studies* 32 (4) (2001) 725-748.

Maes, P., Guttman, R.H., and Moukas, A.G., Agents that Buy and Sell, *Communications of the ACM* 42 (3) (1999) 81-91.

Montgomery, A.L., Hosanagar, K., Krishnan, R., and Clay, K.B., Designing a Better Shopbot, *Management Science* 50 (2) (2004) 189-206.

O'Keefe, R.M. and McEachern, T., Web-based Customer Decision Support System, *Communications of ACM* 41 (3) (1998) 71-78.

Resnik, P., Using Information Content to Evaluate Semantic Similarity in a Taxonomy, *Proceedings of the 14th International Joint Conference of Artificial Intelligence*, Montreal, Canada, 1995, pp. 448-453.

Stumme, G. and Maedche, A., FCA-MERGE: Bottom-up Merging of Ontologies, *Proceedings of the 17th International Conference on Artificial Intelligence*, Seattle, USA, 2001, pp. 225-230.

Tho, Q.-T., Hu, S.-C., Fong, A., and Cao, T.-H., Automatic Fuzzy Ontology Generation for Semantic Web, *IEEE Computer Society* 18 (6) (2006) 842-856.

Turban, E., King, D., McKay, J., Marshall, P., Lee, J., and Viehland, D., *Electronic Commerce: A Managerial Perspective* (5 ed.), Prentice Hall, New Jersey, 2008.

Weng, S.-S., Tsai, H.-J., Liu, S.-C., and Hsu, C.-H., Ontology Construction for Information Classification, *Expert Systems with Application* 31 (1) (2006) 1-12.

Wille, R., Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts, in: I. Rival, Ed., *Ordered Sets*, pp. 455-470 (Reidel, Dordrecht-Boston, 1982).