A Commodity Search System for Online Shopping Based on Ontology and Web Mining

Ming-Hsiung Ying*, Yeh-Yen Hsu†
Department of Information Management, Chung Hua University, Hsinchu, Taiwan
*mhying@chu.edu.tw, †m10110001@chu.edu.tw

Keywords: Commodity Search System , Web Mining; Ontology; Semantic Analysis; Agent

Abstract

With the popularity of Internet and e-commerce, the number of shopping websites has rapidly increased on the Internet, and this enables people to shop easily through the Internet. Consumers spend a lot of time searching commodity, because they need to filter and compare search results data by themselves. In recent years, there is a growing parity websites helping consumers to buy cheaper commodity. Although these websites can help consumers get the parity price of commodities, the search results are not so ideal. Because these websites may occur problems about the difference commodity between search results and consumers want to search, or the difference commodity price between search results and commodity web page. Therefore, this study attempts to use semantic analysis, ontology, and web mining technique as a basic approach. This study proposes a novel commodity search system to track consumer demand, and that is, when the commodity price of any website is lower than the consumer price conditions, the system will proactively notify consumers. This study results indicate that the novel commodity search system could assist consumers to search commodity, and provide historical price information of commodity for consumers to decide.

1 INTRODUCTION

With the Internet becoming more universal, e-commerce gradually on the rise, more and more enterprises are showing up, businesses have a stores entities and creating virtual store that is shopping websites on the internet. In addition, there are many shopping platforms that provide businesses or individuals to sell commodities on the internet.

Now more and more people use shopping websites mainly because shopping websites can provide commodity that are cheaper than those in the physical stores. When consumers search for their desired commodity on so many shopping websites, they need to filter and compare search results and compare different price on different shopping websites by themselves. Therefore, it always takes lots of time for the consumers, and even the search results do not accord with consumers

demand.

With the gradual emergence of many parity websites during these years, parity websites can help consumers to search commodity and a fare price in particular shopping websites. Although parity websites do help, sometimes the difference commodity between search results and consumers want to search, or the difference commodity price information between search results and commodity web page.

Current shopping websites and parity websites won't record commodities search conditions when search results and consumer demand do not accord. In the future, shopping websites and parity websites will not take the initiative instead of consumers to shopping websites search commodity of demand. Therefore, if consumers want to search commodity of demand, they need to search commodity of demand on shopping websites again by themselves, which is very time wasting. Although some shopping websites provide replenishment notification for commodity, but it is still unable to solve time difference commodity between sale and sale off lead to problem that is consumers unable to search commodity for consumer demand do not accord.

In order to solve the above problems, this study attempts to use Semantic Analysis, Ontology, and Web Mining Technique as a basic approach. This study will analyze commodity prices, and provide the relevant information to consumers' reference. This study system will aid consumers to search commodity, long term track for consumer demand, and proactively notify track results.

2 LITERATURE REVIEW

2.1 Ontology

The earliest ontology proposed by Bunge (1977) in the field of philosophy describes the presence and the relationship of things in the real world, investigating the entity of existence, and a description of systematic(Smith & Welty, 2001). Ontology is the basic terms and relations that include subject domain vocabulary and rules for combination words and define vocabulary relationship of expansion (Neches et al., 1991).

Chandrasekaran et al. (1999) indicate be able to understand the structure of the domain knowledge when ontology analysis applied to specific domains, because ontology can knowledge conceptualization, and description relationship between the concepts. Huang (2003) indicates web page and resource should have a knowledge ontology of their own definition that is describe and define any one web page, resources knowledge content, and information architecture.

Authorized licensed use limited to: Xiamen University. Downloaded on May 18,2022 at 08:44:17 UTC from IEEE Xplore. Restrictions apply.

2.2 Web mining

The earliest Web Mining is data mining technology combing Web technology used on the Internet by Etzioni (1996), and the discovery of useful information to be analyzed (Caverlee, 2004; Nie & Kambhampati, 2004; Cooley et al., 1997). Web mining is defined as the use of data mining technology for internet files and services to find and extract the hidden information by Etzioni.

2.3 Semantic web

Tim Berners-Lee (1999) proposed World Wide Web that is Universal Resource Identifier, Hypertext Transform Protocol, and Hypertext Markup Language composition. The information through hyperlinks series that is only people can interpret, but the computer does not understand the meaning (Chu-Ren Huang et al., 2003).

The study of Chu-Ren Huang et al. (2003) indicates must complete expression the phenomenon of particular domain to let the computer understand the semantic, but maybe these have different words to express the same kind of phenomenon. Therefore, we must understand single relationship between language vocabularies and correspond with convert between languages.

2.4 Chinese word segmentation

The process of Chinese word segmentation is a very important step for Chinese string process (Qian-Xiang Lin et al., 2010).Due to syntactic and semantic of Chinese basic units are vocabulary instead word that are not be separated between each of Chinese word by any symbol lead more difficult to judge. The syntax and semantics of English basic unit word that are be separated between each of English word, and therefore easy to judge (Jiah-Shing Chen et al., 2000).

In Taiwan, Chinese Knowledge Information Processing Group (CKIP) are develop Chinese Word auto Segmentation System in order to be able to process Chinese word segmentation. The basic principle of this system is construct Sinica Treebank. When natural language of Chinese processing, Sinica Treebank can provide an important marker corpus, and extract grammatical knowledge from Sinica Treebank database for Chinese analysis system become more complete and accurate (Feng-Yi Chen et al.,1999). When Chinese string through CKIP's Chinese Word auto Segmentation System processing, it will automatically speech marking for vocabulary of the results.

3 Research Methods and Design

3.1 Commodity information ontology concept design

This study is to make Commodity Search System to understand what information is contained in the commodity, what is more important to consumers, and what belongs to a specific category for each commodity.

Therefore, this study will refer to Taiwan's major shopping websites in the commodity information, analyze attributes of commodities, and find specific attributes of the same type of commodities will be classified, while presenting a commodity information ontology in Fig. 1.

This study proposes commodity information ontology, including name of commodity, brand, category, manufacturer, made in, and attributes of each category.

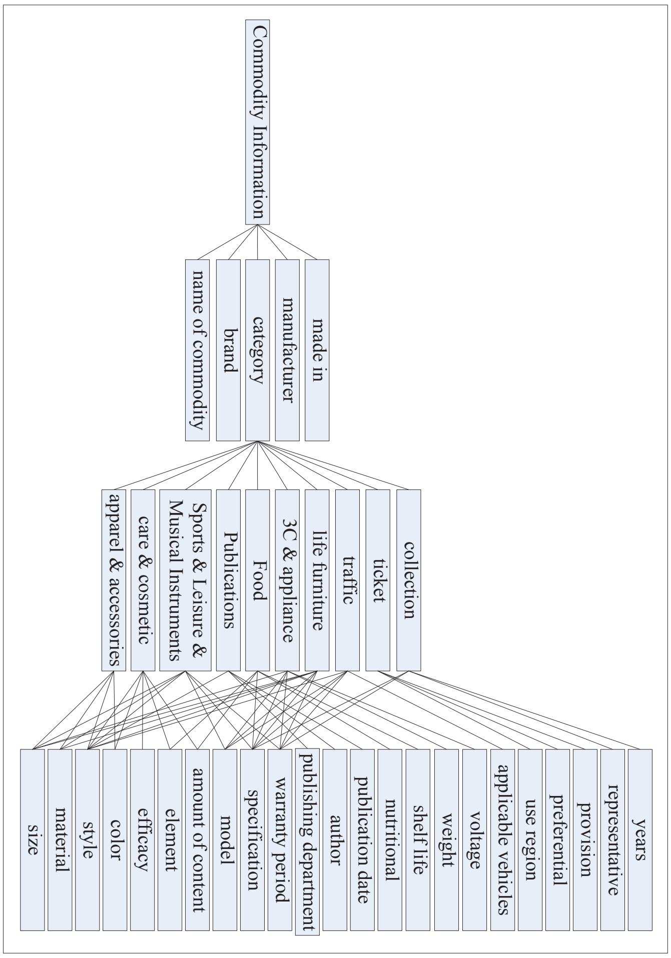


Figure 1: Commodity Information Ontology

3.2 Commodities information crawl agent

Consumers search commodity takes a lot of time on different commodities of shopping websites between searching and comparing. In order to reduce the time costs for consumers, this study designs commodities information crawl agent that is through computer and time schedule to crawl commodities data of websites on Web Ming Technology. Their processes are described as follows in Fig.2.

- The first reads the URL of target shopping websites from shopping websites source database.
- The second demand is sent to the target shopping website server and download commodities content HTML of the target shopping websites.
- The third obtain commodities subject relevant information from commodities content HTML, and determine whether commodity already exist in the commodities source database.

Authorized licensed use limited to: Xiamen University. Downloaded on May 18,2022 at 08:44:17 UTC from IEEE Xplore. Restrictions apply.

- If judgment is yes that is update commodity price and add commodity price history in the database, otherwise store commodity data in the database.
- Last determine whether reading the finished shopping websites sources, if judgment is yes that is commodities data module process commodities data, otherwise to continue reads the shopping website sources

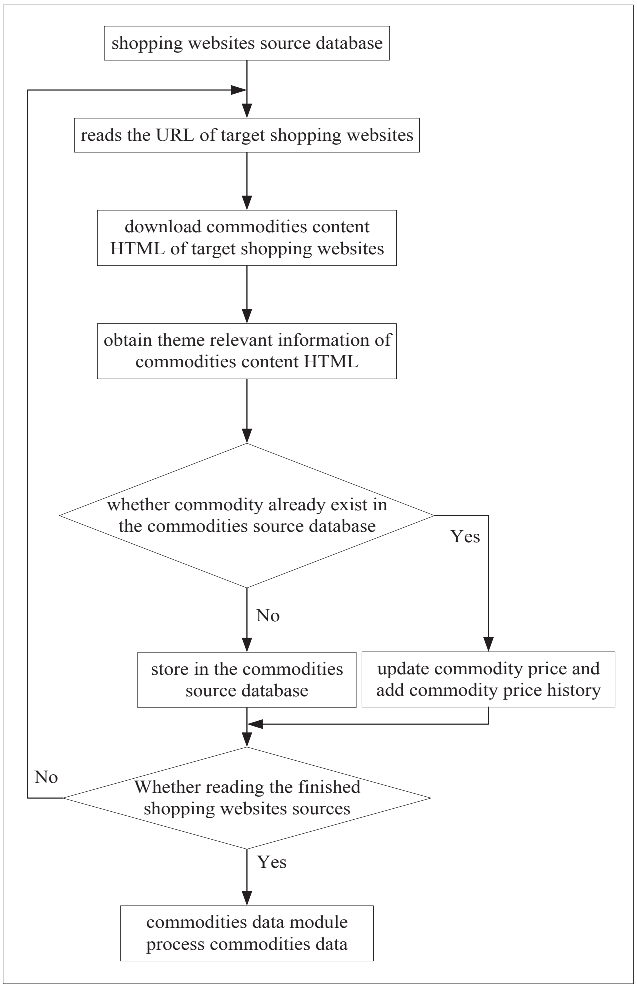

shopping websites source database
reads the URL of target shopping websites
download commodities content HTML of target shopping websites
obtain theme relevant information of commodities content HTML
whether commodity already exist in the commodities source database
Yes
No
store in the commodities source database
update commodity price and add commodity price history
No
Whether reading the finished shopping websites sources
Yes
commodities data module process commodities data

Figure 2: Commodities information crawl agent process

3.3 Commodities search agent

When consumers search commodities, sometimes the search results and consumer demand do not accord. Therefore, this study designs commodities search agent that accords with consumer search demand, and regular replace consumers for search commodity in the shopping website. Their processes are described as follows in Fig.3.

- The first reads consumers search conditions from demand conditions aggregated module and the URL of target shopping websites from shopping websites source database.
- The second demand is sent to the target shopping website server and download commodities content HTML of the target shopping websites.
- The third determine whether commodity is exist, if judgment is yes that is obtain commodities subject relevant information from commodities content HTML, otherwise determine whether reading the finished shopping websites sources.
- The fourth determine whether commodity already exist in the commodities source database, if judgment is yes that is update commodity price and add commodity price history in the database, otherwise store commodity data in the database.
- Last determine whether reading the finished shopping websites sources, if judgment is yes that is determine whether reading the finished search conditions, otherwise to continue reads the shopping website sources.
- If judgment is yes that is commodities data module process commodities data, otherwise continue reads search conditions.

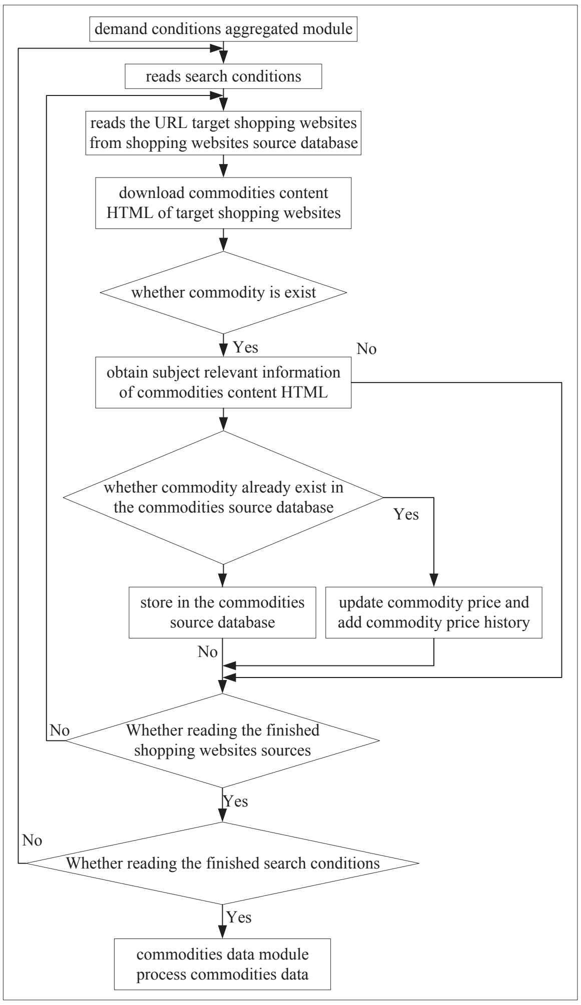

demand conditions aggregated module
reads search conditions
reads the URL target shopping websites from shopping websites source database
download commodities content HTML of target shopping websites
whether commodity is exist
Yes
No
obtain subject relevant information of commodities content HTML
whether commodity already exist in the commodities source database
Yes
store in the commodities source database
update commodity price and add commodity price history
No
No
Whether reading the finished shopping websites sources
Yes
No
Whether reading the finished search conditions
Yes
commodities data module process commodities data

Figure 3: Commodities search agent process

Authorized licensed use limited to: Xiamen University. Downloaded on May 18,2022 at 08:44:17 UTC from IEEE Xplore. Restrictions apply.

3.4 Tracking commodities compare agent

This study designs tracking commodities compare agent that compares demand conditions of consumers and commodities data, and initiative to inform consumers. Their processes are described as follows in Fig.4.

- The first reads demand conditions of members from demand conditions database, and compare commodities from commodities database.
- The second determine whether commodity is meet, if judgment is yes that is reads source data of commodities from commodities source database and reads contacts information of members from members database, otherwise to continue reads demand conditions of members.
- Last determine whether processing the finished demand conditions, if judgment is yes that is the results of meet to inform members, otherwise to continue reads demand conditions of members.

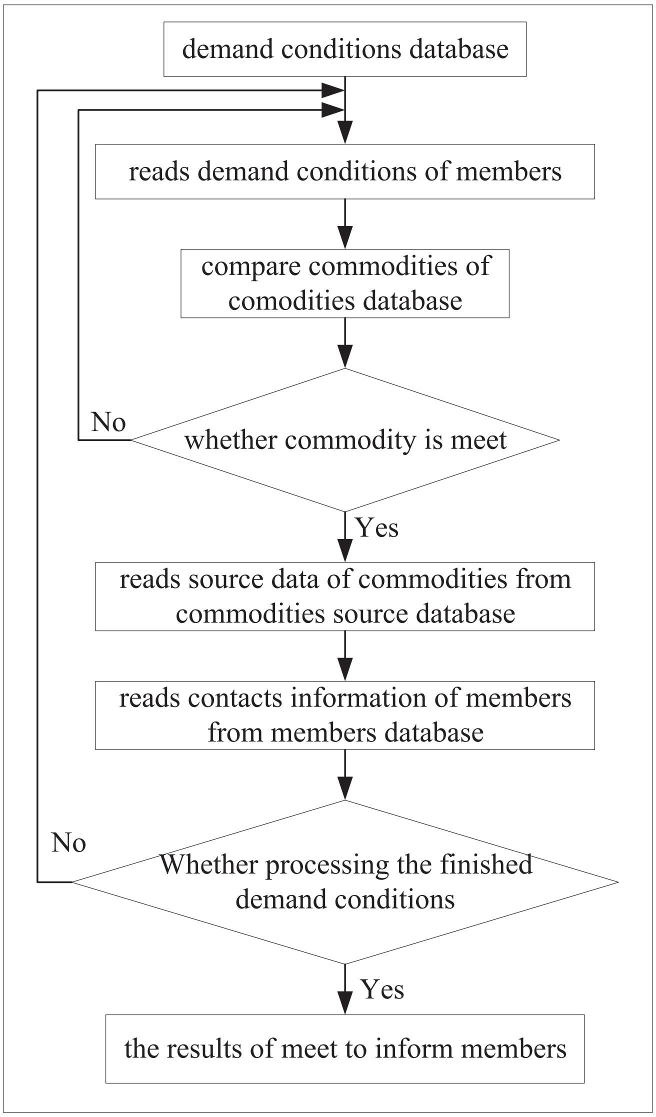

demand conditions database
reads demand conditions of members
compare commodities of comodities database
No
whether commodity is meet
Yes
reads source data of commodities from commodities source database
reads contacts information of members from members database
No
Whether processing the finished demand conditions
Yes
the results of meet to inform members

Figure 4: Tracking commodities compare agent process

3.5 Semantic rules build and correspond

This study Chinese word processing is the use CKIP system of Academia Sinica, through CKIP process will be given speech tags each vocabulary, but do not represent specific domain. Therefore, this study was the analysis of commodity related data processing that is will be through vocabulary of domain repair then give the corresponding of speech by speech tags custom tables in this study in Table1. This allows vocabulary gives a specific meaning in this domain, so that subsequent processing easier.

Table 1: The speech tags custom tables

tags of speech	vocabulary example	represent of the meaning and description
A	X552VL, 0A36283 ⋯	represent special vocabulary (such as model)
B	Asus, Acer⋯	represent brand
C	系列、長效⋯ (series,long acting..)	represent no special meaning
DM	5g, 300ml ⋯	represent amount of content
Symbol	. , & ⋯	represent symbol
⋮	⋮	⋮

In this study, the specific attribute combinations of commodity category to build rules for the system is easy to determine commodity data. In this category of apparel & accessories by size, material, style, color, composition for example in Table2.

Table 2: The rules of apparel & accessories

number	sample content rules
1	…size (C1)…material (C2)…style (C3)…color (C4)…
2	…size (C1)…material (C2)…style (C3)…
3	…size (C1)…style (C3)…
⋮	⋮

3.6 Commodity subject analysis

The Commodities subject is by name of commodity, features, specifications and marketing vocabulary composed in the most shopping websites. In order to integrate the same commodities together, we must extract more realistic name of commodity. Therefore, this study proposes a name of commodity processing program and verify Equation, as in (1), (2).

This study will be the commodities subject through string processing, compare name of commodity, and using (1). When

Authorized licensed use limited to: Xiamen University. Downloaded on May 18,2022 at 08:44:17 UTC from IEEE Xplore. Restrictions apply.

a similar rate, the higher the more likely for the same commodity. In this study, a similar rate of greater than 0.8 is set to true name of commodity.

When compare commodities do not accord with the commodities database, then different shopping websites will search and compare the similarities rate between commodities subject, and using (2) to obtain up of a combination of two similar rate.

$$A_{(J)} = \frac{SL}{\sqrt{SL} \times \sqrt{L_{(j)}}} \qquad (1)$$

- j: Represents the j is name of commodity of query results in the commodities database, j=1,2,...,m, m is the total number of commodities.
- $L_{(j)}$: Represents the j is name length of commodity of query results.
- SL: Represents meet the total length of the set of query vocabulary.
- $A_{(j)}$: Represents similar rates of the j is name of commodity vocabulary sets and query results Meet the query.

$$B_{(z)} = \frac{WN_{(z)}}{\sqrt{PN} \times \sqrt{RN_{(z)}}} \qquad (2)$$

- $B_{(z)}$: Represents similar rates of the z is vocabulary sets in the search vocabulary sets.
- RN(z): Represents the total number of the z is vocabulary sets in the search results of commodities subject vocabulary sets
- PN: Represents the total number of primary commodities subject vocabulary sets
- WN(z): Represents the total number of the z is vocabulary sets in the vocabulary sets of search.

3.7 Commodity prices reasonable range analysis

Due to commodity prices of shopping websites are likely errors or commodities specials, rather than causing too much difference between the prices of the same commodities. Therefore, this study process prices of commodities using (3), (4), (5), and three standard deviation is a reasonable range. When the price exceeds the reasonable range, using 5% of the average price of spread for the new a reasonable range, and tags might be specials.

$$DP_{(i)} = \frac{1}{m} \times \sum_{j=1}^{m} P_{ij} \quad (3)$$

- i: Represents the i day, i=1,2,…,n, n is total days.
- j: Represents the j commodity source of the i day, j=1,2,…,m, m is total commodities.
- P_{ij}: Represents the j commodity source price of the i day.
- $DP_{(i)}$: Represents the average price of the i day.

$$x = \frac{1}{n}\sum_{i=1}^{n} DP_{(i)} \qquad (4)$$

- x: Represents the average price of sample (n days).

$$\sigma = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n}\left(DP_{(i)} - x\right)^2} \qquad (5)$$

- σ : Standard deviation

4 System Implementation and Show

4.1 Commodities search system environment schema

In this study, the system agent retrieves data and commodities data processing purposes. Therefore, division of labor with two servers to reduce the time to retrieve data and reduce the burden on the server in Fig.5. Master server provides users client services, and the sub server auxiliary master server. Master server will assign commands to give sub server, and then sub server execute the task order.

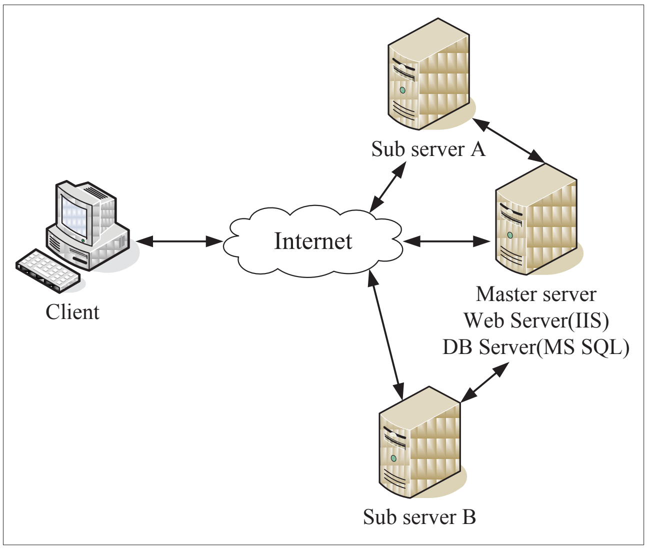


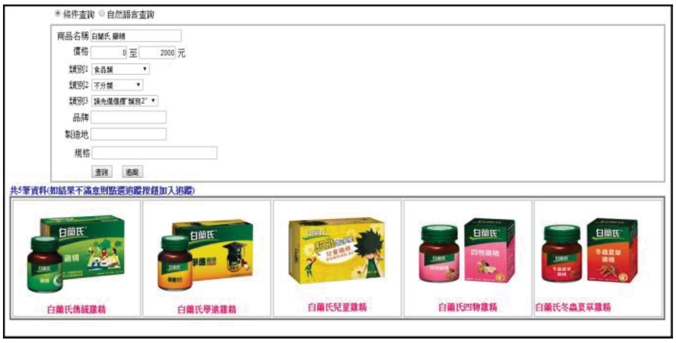
Figure 5: Commodities search system environment

4.2 Commodities search and track

In this study, consumers can search or track conditions according to demand commodities in Fig.6, Fig.7.

If search result is dissatisfied then the system will record consumer demand conditions, and replace to consumers for regular long term tracking.

Authorized licensed use limited to: Xiamen University. Downloaded on May 18,2022 at 08:44:17 UTC from IEEE Xplore. Restrictions apply.



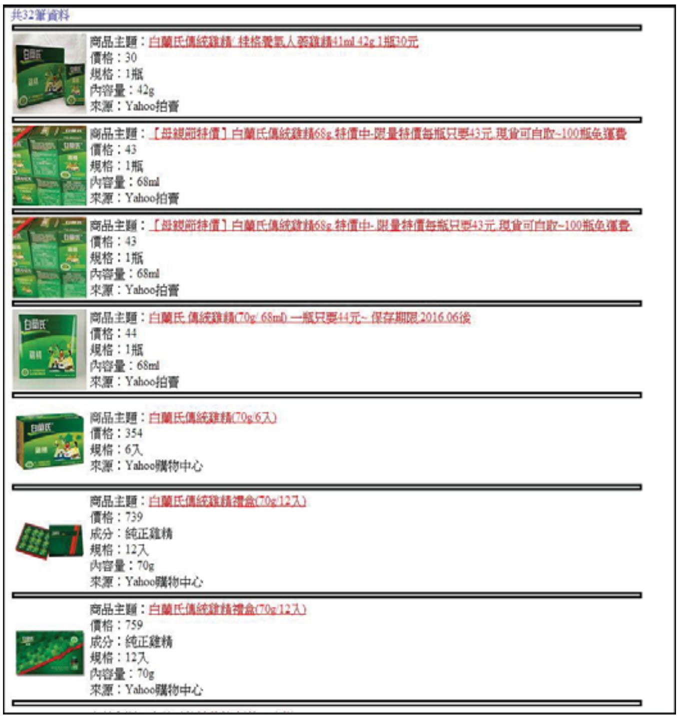
● 條件查詢 ○ 自然語言查詢

商品名稱 白蘭氏 雞精
價格 0 至 2000 元
類別1 食品類 ▾
類別2 不分類 ▾
類別3 請先選擇"類別2" ▾
品牌
製造地
規格
查詢 追蹤

共5筆資料(如結果不滿意則點選追蹤按鈕加入追蹤)

白蘭氏傳統雞精
白蘭氏學進雞精
白蘭氏兒童雞精
白蘭氏四物雞精
白蘭氏冬蟲夏草雞精

Figure 6: User search and track screen



共32筆資料

商品主題：白蘭氏傳統雞精/ 桂格養氣人蔘雞精41ml 42g 1瓶30元
價格：30
規格：1瓶
內容量：42g
來源：Yahoo拍賣

商品主題：【母親節特價】白蘭氏傳統雞精68g 特價中-限量特價每瓶只要43元 現貨可自取~100瓶免運費
價格：43
規格：1瓶
內容量：68ml
來源：Yahoo拍賣

商品主題：【母親節特價】白蘭氏傳統雞精68g 特價中-限量特價每瓶只要43元 現貨可自取~100瓶免運費
價格：43
規格：1瓶
內容量：68ml
來源：Yahoo拍賣

商品主題：白蘭氏 傳統雞精(70g/ 68ml) 一瓶只要44元~ 保存期限2016.06後
價格：44
規格：1瓶
內容量：68ml
來源：Yahoo拍賣

商品主題：白蘭氏傳統雞精(70g/6入)
價格：354
規格：6入
來源：Yahoo購物中心

商品主題：白蘭氏傳統雞精禮盒(70g/12入)
價格：739
成分：純正雞精
規格：12入
內容量：70g
來源：Yahoo購物中心

商品主題：白蘭氏傳統雞精禮盒(70g/12入)
價格：759
成分：純正雞精
規格：12入
內容量：70g
來源：Yahoo購物中心

Figure 7: Commodities source data appear

5 CONCLUSION

This study designed three different uses of the agent to aid in searching commodities. The commodities information crawl agent will download commodities saved in the database, so that consumers can search commodities on this study system. If result is not satisfactory can be used tracking feature, commodities search agent take the initiative to regular replace consumers for search commodity in the shopping website, and tracking commodities compare agent compare demand conditions of consumers and commodities database of data, and initiative to inform consumers.

The study based on commodity information ontology for categories of each commodities, similar rates are based on commodities subject analysis than the same commodities, analysis commodity prices and historical prices, and providing commodity price volatility and reasonable price range information to consumers.

References

[1] Bunge, M. (1977), “Ontology I: The Furniture of the World. Treaties on Basic Philosophy”, Vol.3, Boston, MA: Reidel.

[2] Smith, B. & Welty, C. (2001), “Ontology: Toward a New Synthesis”, Proceedings of the international conference on Formal Ontology in Information Systems, Ogunquit, Maine, USA.

[3] Neches, R. Fikes, R. Finin, T.Gruber, T., Patil, R., Senator, T. Swartout, W. R. (1991), “Enabling Technology for Knowledge Sharing”, AI Magazine, 12(3), pp.36-56.

[4] Huang, C.R. (2003) “Semantic web, WordNet and Ontology: A talk on knowledge management on future’s web” Information Management for Buddhist Libraries (33), pp.1-16.

[5] Etzioni, O. (1996), “The World Wide Web: Quagmire or Gold Mine”, Communications of the ACM, (39), pp.65-68.

[6] Caverlee, J., Liu, L., & Buttler, D. (2004), “Probe, Cluster, and Discover: Focused Extraction of QA-Pagelets from the Deep Web”, Proceedings of IEEE International Conference on Data Engineering, pp.103-114.

[7] Nie, Z. & Kambhampati, S. (2004), “A Frequency-based Approach for Mining Coverage Statistics in Data Integration”, Proceedings of International Conference on Data Engineering, pp.387-398.

[8] Cooley, R., Mobasher, B., Srivastava, J. (1997), “Web Mining : Information and Pattern Discovery on the World Wide Web ” , Proceedings of Ninth IEEE International Conference of Tools with Artificial Intelligence, pp.558-567.

[9] Berners Lee, T. (1999) “Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web” London: Verso.

[10] Huang, C. R., Chang, R. Y., & Tsai, B. S. (2003) “Chinese Language Education and the Developing Semantic Web: An Introduction to Chinese-English Bilingual Ontology Interface” The Third International Conference on Internet Chinese Education, pp.24-56.

[11] Lin, Q. X., Chang, C. H., & Chen, C. L. (2010) “A Simple and Effective Closed Test for Chinese Word Segmentation Based on Sequence Labeling” Computational Linguistics and Chinese Language Processing, Vol. 15, No. 3-4, pp.161.-180.

[12] Chen, J.S., Hsieh C.L., & Hsu, F.C. (2000) “A Study on Chinese Word Segmentation: Genetic Algorithms Approach” Journal of E-business, 2(2), pp. 27-44.

[13] Chen, F. Y., Tsai, P. F., Chen, K. J., & Huang, C. R. (1999) “Sinica Treebank” Computational Linguistics and Chinese Language Processing, Vol. 4, No. 2, pp.87.-104.

Authorized licensed use limited to: Xiamen University. Downloaded on May 18,2022 at 08:44:17 UTC from IEEE Xplore. Restrictions apply.