An Enhanced Holistic Information Retrieval System

Kwang Mong Sim and Paul C. K. Kwok

Abstract—**This paper presents an enhanced holistic information retrieval (IR) system that aims to automate the entire process of Web-based IR. The system consists of three types of agents: 1) ontology-enhanced Web browsing agents (WBAs) that are used to autonomously browse and scan multiple Websites to determine and rate the relevance of Websites, 2) Web monitoring agents (WMAs) that are used for tracking and reporting changes in selected Websites, and 3) price watcher agents (PWAs) that monitor product prices from competing suppliers' Websites. WBAs perform information filtering by considering three relevance metrics: ontological relations, frequency, and nearness of keywords. The general idea of Website monitoring is that each WMA is programmed to download a new copy of a Website and compare it with the old copy. WMAs allow users to specify monitoring rules, and provide user interface for specifying patterns and data to be monitored. PWAs invoke the functionalities of WBAs and WMAs for browsing and monitoring multiple Websites displaying different prices of a product. Whereas empirical results show that WBAs are likely to rate the relevance of Website with a small degree of error, proof-of-concept examples demonstrate the major functionalities of WMAs and PWAs.**

Index Terms—**Web information retrieval, software agent.**

I. Introduction

Web users searching for information are often overwhelmed with very large numbers of URLs returned from search engines. Whereas many of the URLs are often quite relevant, it is not uncommon that irrelevant Websites containing query keywords are among the suggested URLs because words can have several meanings (senses). For example, a typical user using Google search to search for Websites about "mountain chain" may find the URL http://www.chainreactioncycles.com/[1], which contains words such as "chain" (as in bicycle chain). Consulting WORDNET [1], "chain" has several senses (meanings). One of the senses of chain refers to "a series of (usually metal) rings or links fitted into one another to make a flexible ligament" and another refers to "a series of hills or mountains". In such situation, one possible solution is to program a software agent to distinguish between relevant and irrelevant URLs by searching for evidence phrases by consulting an ontology [2]. Evidence phrases may include ontologically related words such as synonym, hyponym, hypernym, meronym, and holonym. For example, Websites containing words such as "Adirondack Mountains" and "Alaska Range " (a hyponym of mountain chain) are more likely to contain relevant information about mountain chain than a Website with words like "anchor chain " and "tire chain" (a hyponym of "iron chain ").

Furthermore, even though users can use search engines to locate URLs and program software agents to autonomously browse selected Website(s), they still need to repeatedly and regularly visit the Websites to retrieve up-to-date information. Due to the ever-changing content of Webpages, tracking the changing contents of Websites may be tedious and time-consuming. Examples of Websites with ever-changing contents include financial Websites that display stock prices and Websites of retail companies that display prices of computer products and accessories. It is not uncommon that investors constantly visit multiple financial Websites and continuously monitor stock prices, and analyze stock trends. One way of assisting such users is to build software tools that visit selected Websites to monitor and track changes in the contents of these Websites. Additionally, software agents for bolstering price comparisons among multiple Websites selling the same product may also be useful tools for retailers and e-shoppers.

The objective of this project is to develop a holistic information retrieval (*IR*) system (section II) that augments the functionalities of existing search engines by supporting the following:

1. Autonomous filtering of contents in Websites,
2. Regularly monitoring and reporting (selected) changes in Websites, and
3. Regularly comparing and reporting product prices from competing suppliers.

This project is designed to support the information gathering activities in Ellis' model [3] that are not supported by existing search engines. Ellis' model [3] of information gathering includes 1) activities that form initial search for information by following and linking to other information sources, 2) browsing (scanning information source), 3) differentiating (filtering and selecting among the sources), 4) monitoring (regularly following a particular source) and 5) extracting (identifying materials of interest from some sources). Whereas activity (1) is supported by existing search engines, the enhanced holistic *IR* system in this project is designed to bolster activities (2) through (5). To this end, this project complements and augments the functionalities of existing search engines. In particular, it is reminded here that

Manuscript received November 27, 2008. This work was supported by the DASAN Project (GIST International Faculty Start-up Fund) from the Gwangju Institute of Science and Technology, South Korea.

Kwang Mong Sim is with the Gwangju Institute of Science and Technology, South Korea (phone: 82-62-970-2235; fax: 82-62-970-2204; e-mail: kmsim@ gist.ac.kr).

Paul C. K. Kwok is with the School of Science and Technology, Open University of Hong Kong (e-mail: ckkwok@ouhk.edu.hk).

[1] This query was executed on Google on November 18, 2008. This may change due to changes in the Website or Google.

this project does not compete with existing search engines and is certainly not designed to replace existing search engines, but rather to supplement their functionalities.

II. A Holistic IR system

This section presents the prototype of an enhanced holistic *IR* system consisting of three types of agents: 1) Web Browsing Agents (*WBAs*), 2) Web Monitoring Agents (*WMAs*), and 3) Price Watcher Agents (*PWAs*).

Web Browsing Agent: A *WBA* supports a user by bolstering activities (2) and (3) of Ellis' model of information gathering (section I). That is, it performs the following tasks: i) browsing and scanning the information contents of a Website and ii) determining the relevance of and rating the contents in a Website. Based on Sim's previous works [4-6], details of the functionalities of a *WBA* are given in section III.

Web Monitoring Agent: A *WMA* supports a user by bolstering activities (4) and (5) of Ellis' model of information gathering (section I). It carries out the following tasks: 1) regularly monitoring a selected Website and tracking changes in the Website, and 2) identifying and reporting selected changes in the contents of the Website that it is monitoring.

Price Watcher Agent: A *PWA* supports a user by invoking a search engine to search for Websites containing the prices of a product, 2) deploying multiple *WBAs* for determining and rating the relevance of a list of Websites displaying that product, 3) deploying multiple *WMAs* for monitoring changes in product prices in multiple Websites, and 4) displaying in ascending order the prices of that product from different Websites.

Stages of Information Gathering: The stages of the information gathering process are listed and described as follows:

1. *Locating Information Resources*. This is typically the first step that a user would do when searching for information through the Web – compose a query using a set of keywords, then enter the query to a search engine.

2. *Browsing and Evaluating Selected Websites*. When a search engine returns a set of URLs, *multiple WBAs* are deployed to *simultaneously* visit and browse the contents of the set of URLs and verify if the contents are relevant to a query. This corresponds to the second step that a user would typically do – visiting, browsing, and deciding if the contents of the URLs are relevant.

3. *Monitoring Changes in Selected Websites*. In this stage, a set of *WMAs* is activated to monitor and track changes in selected relevant URLs. This stage corresponds to a user bookmarking a set of favorite URLs and perhaps repeatedly visiting the URL to retrieve updated and ever-changing information (e.g., stock prices).

III. Web Browsing Agent

A *WBA* carries out two functions : 1) information filtering and 2) information rating.

Information Filtering : In this stage, a *WBA* adopts WORDNET's ontology [1] for determining the relevance of a Website. This is achieved by constructing a set of evidence phrases for a user query by considering ontological relations from WORDNET such as *meronym* and *holonym*. Meronym and holonym refer to the part-whole relations of words [7]. Whereas a meronym is the name of a constituent part of a concept, a holonym is the name of the whole of which the meronym is a part (i.e., P_1 is a meronym of Q_1 if P_1 is a part of Q_1, and Q_2 is a holonym of P_2 if P_2 is a part of Q_2) [1]. Furthermore, in WORDNET, some of the categories of meronym relations include :

1) part mernonym : P_1 is a *part meronym* of Q_1 if P_1 is a *component part* of Q_1.
Example Query Word: battery
Part Meronym(s) : electrode, pole (terminal).

2) member meronym : P_1 is *a member meronym* of Q_1 if P_1 is a *member* of Q_1
Example Query Word : Forest
Member Meronym(s) : tree

3) substance meronym: P_1 is a *substance meronym* of Q_1 if P_1 is the *stuff that Q_1 is made of.*
Example Query Word : chalk
Substance Meronym(s) : calcium carbonate

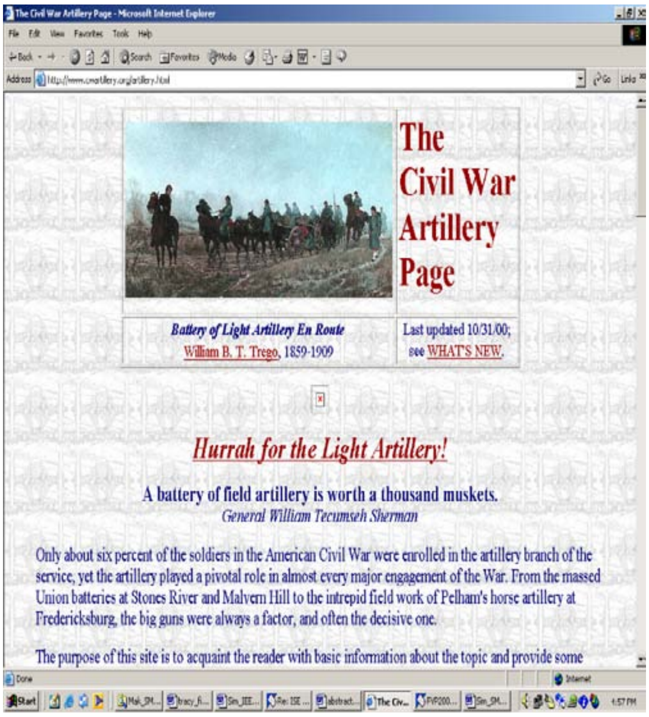
When filtering relevant URLs, a *WBA* examines the content of a URL for meronyms of a query word by consulting *WORDNET*. Given that each word can have different senses, an irrelevant URL is identified by searching for meronyms of query keywords of other senses. For instance, to identify an irrelevant Webpage for the query "battery" (in the sense of "electric battery"), the *WBA* filters out Webpages with meronyms such as "gun" and "missile launcher" which are meronyms of battery in the sense of gunnery. An example of an irrelevant Webpage containing the keyword "battery" for the query "battery" (in the sense of "electric battery") is shown in Fig.1. Fig. 1 shows a Webpage of "battery" in sense of gunnery. Such an irrelevant Webpage can be identified when a *WBA* searches for phrases such as "missile launcher" and "gun". To verify that a Webpage is relevant to the query "battery" in the sense of "electric battery", a *WBA* should search for evidence phrases such as "electrode" and "pole" (meronyms of electric battery). If a Webpage is determined to be relevant, a *WBA* proceeds to rate the Webpage.

Information Rating : A *WBA* rates the information contents in a Website by considering 3 heuristic factors: 1) ontologically related words, 2) frequency of occurrence, and 3) nearness of keywords.

(1) By searching for ontologically related words in a Webpage, a *WBA* is more likely to detect information related to a query. To ensure high precision, the heuristic in [5, p. 96] is used to guide a *WBA* in identifying relevant information of different degrees.

(2) The relevance metric used by a *WBA* favors Webpages with higher occurrence of keywords in a user query. For example, if the term "car" occurred reasonably frequently in a Webpage, it seems plausible to think that the Webpage contains information that deals with "car" [8, pp. 279-280].

(3) When nearness [8, pp. 237] is included in the relevance metric, the probable relevance of the information retrieved is likely to be higher. For instance, consider the query "nature picture", if both "nature" and "picture" occur adjacently in a given Webpage, then it is more likely that the Webpage contains more relevant information than when both "nature" and "picture" occur within a sentence but are separated by some words".


The Civil War Artillery Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.cwartillery.org/artillery.html
Go Links
Done
Internet
Start
4:57 PM

The Civil War Artillery Page

Battery of Light Artillery En Route
William B. T. Trego, 1859-1909

Last updated 10/31/00; see WHAT'S NEW.

Hurrah for the Light Artillery!

A battery of field artillery is worth a thousand muskets.
General William Tecumseh Sherman

Only about six percent of the soldiers in the American Civil War were enrolled in the artillery branch of the service, yet the artillery played a pivotal role in almost every major engagement of the War. From the massed Union batteries at Stones River and Malvern Hill to the intrepid field work of Pelham's horse artillery at Fredericksburg, the big guns were always a factor, and often the decisive one.

The purpose of this site is to acquaint the reader with basic information about the topic and provide some

Fig. 1. An Irrelevant Webpage for a Query Search

IV. Web Monitoring Agent

A *WMA* carries out two functions : 1) monitors changes in a Webpage, and 2) extracts specific information from a Webpage periodically. The general idea of information monitoring is to download a new copy of a Website and compare it with the old copy. An algorithm for monitoring changes in a table within a Website is given in algorithm 1 [4,6].

Algorithm 1:

1. Retrieve the Webpage from the given URL at time t as a string of characters $S(t)$.
2. Extract all tables from $S(t)$ as $\{T_1,T_2, \ldots, T_n\}$. For each T_x in $S(t)$, T_x contains a set of cells $\{c_{1,1},c_{1,2},\ldots, c_{r,c}\}$
3. Let Value($c_{x,y}$,t) return the value of $c_{x,y}$ at time t and Type($c_{x,y}$, t) be the data type of the $c_{x,y}$ at time t that can either be a string or numeric type. If Value($c_{x,y}$, t) only contains digits , period(".") and comma(","), it is assumed that Value($c_{x,y}$, t) is numeric. All other values are considered as string.
4. Select a cell, $c_{x,y}$,from $\{c_{1,1},c_{1,2},c_{1,3},\ldots\}$ to monitor changes.
5. If the Type($c_{x,y}$, t) is string, report changes if Value ($c_{x,y}$, t_{n+1}) is different from Value ($c_{x,y}$, t_n). This is accomplished by comparing the strings at t_n and t_{n+1}.
6. If the Type($c_{x,y}$, t) is numeric, report changes if

$$\left|\frac{\text{Value}\,(c_{x,y}\,,t_{n+1}) - \text{Value}\,(c_{x,y}\,,t_n)}{\text{Value}\,(c_{x,y}\,,t_n)}\right| > n\%$$

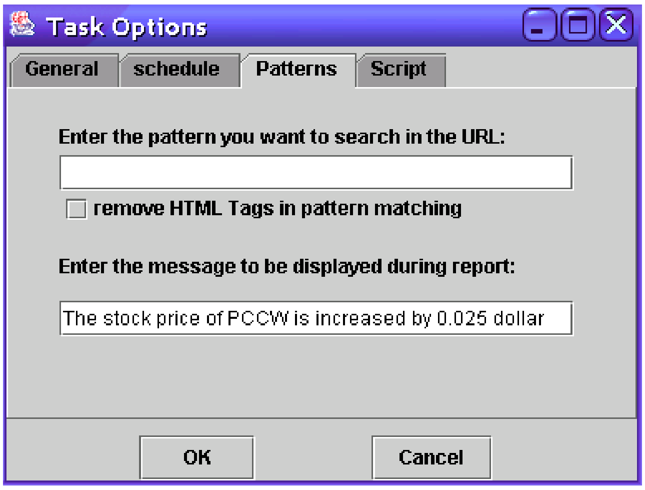
where n is a user defined threshold

Monitoring rules : In a *WMA*, each monitoring task can be represented by a task script using monitoring rules. Some of the monitoring rules in a *WMA* are given as follows :

R1 : If (modified()) notify();
R2 : If (modified()) download();
R3 : If (new(2)>2.0) notify();
R4 : If (new(2)-old(2)>=0.1) notify()

R1 (respectively, *R2*) simply specifies that the *WBA* should notify a user (e.g., by sending email) (respectively, download the Webpage to local disk instead of notifying the user) if there is any changes in a Webpage. Used in conjunction with a pattern, the functions old() and new() in *R3* and *R4* refer to the old and new values of the data associated to a pattern. For instance, if a user instructs a *WMA* to monitor the value of a stock called "tom.com" (see section VI) then in *R3* and *R4*, "2" inside the functions old() and new() is a marker that points to a data value associated to the pattern "tom.com", i.e., "2" refers to the stock price of "tom.com" displayed in the Webpage that the *WMA* is monitoring.

Whereas regular expression is used for specifying the instructions for a *WMA* to extract specific information from a Website, a user interface is developed to allow a user to specify : 1) the pattern that a *WMA* should monitor and the message it should display (Fig. 2), 2) the interval for monitoring the Webpage (Fig. 3), and 3) the position of the data that is associated with a pattern (Fig. 4).



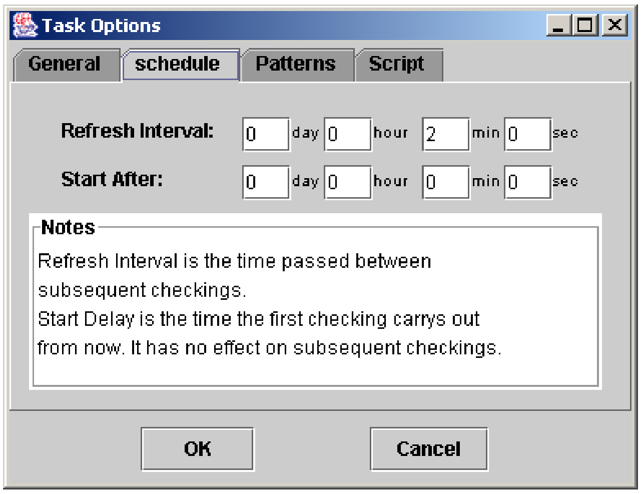
Enter the pattern you want to search in the URL:

□ remove HTML Tags in pattern matching

Enter the message to be displayed during report:

The stock price of PCCW is increased by 0.025 dollar

Fig. 2. *WMA* User Interface : Pattern to Monitor



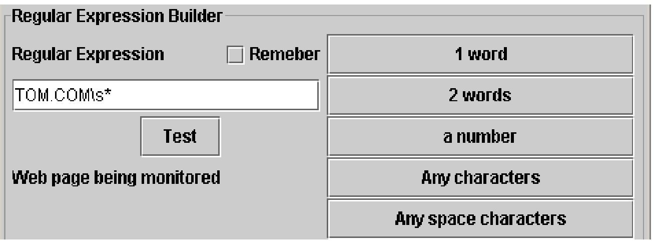
Refresh Interval: 0 day 0 hour 2 min 0 sec

Start After: 0 day 0 hour 0 min 0 sec

Notes

Refresh Interval is the time passed between subsequent checkings.
Start Delay is the time the first checking carrys out from now. It has no effect on subsequent checkings.

Fig. 3. *WMA* User Interface : Schedule to Monitor.



Regular Expression: TOM.COM\s*

Web page being monitored

Fig. 4. *WMA* Pattern Builder Interface.

V. Experimentation and Evaluation

Evaluation of the *WBA* consisted of 1) user study, and 2) a series of experiments using the *WBA* for rating the same set of URLs rated by human users. For user study, two human users were asked to rate the relevance of the top 5 URLs returned by Google for 100 queries. Using a *WBA* to rate the relevance of contents in URLs, two series of experiments were conducted [4-6]. For the first series of experiments, the *WBA* was programmed to *incrementally* recognize exact words, synonyms, hyponyms, and hypernyms (see Table I). The second series of experiments examined the effect of using five different combinations of weightings of the three heuristics in section III: 1) ontological related words (*OR*), 2) frequency of occurrence (*FO*), and 3) nearness of keywords (*NK*) (see Table II).

Table I. *WBA* Combination of Related Terms

Simulation	Word relations
WBA1	{Exact words}
WBA2	{Exact words} + {synonyms}
WBA3	{Exact words, synonyms} + {hyponyms}
WBA4	{Exact words, synonyms, hyponyms} + {hypernyms}

Table II. Weightings of the 3 Heuristics

Weight combination	*OR*	*FO*	*NK*	Difference between users and *WBA* ratings
1	0.6	0.2	0.2	29%
2	0.34	0.33	0.33	16%
3	0.5	0.25	0.25	21%
4	0.6	0.3	0.1	29%
5	0.4	0.4	0.3	18%

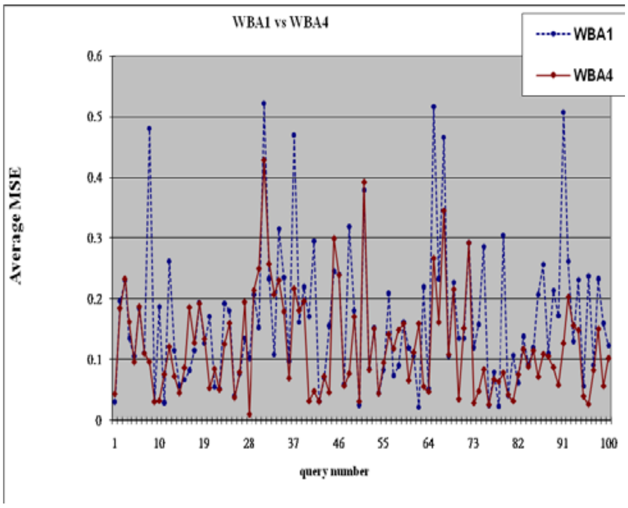


Fig. 5. Differences in User and *WBA* Ratings

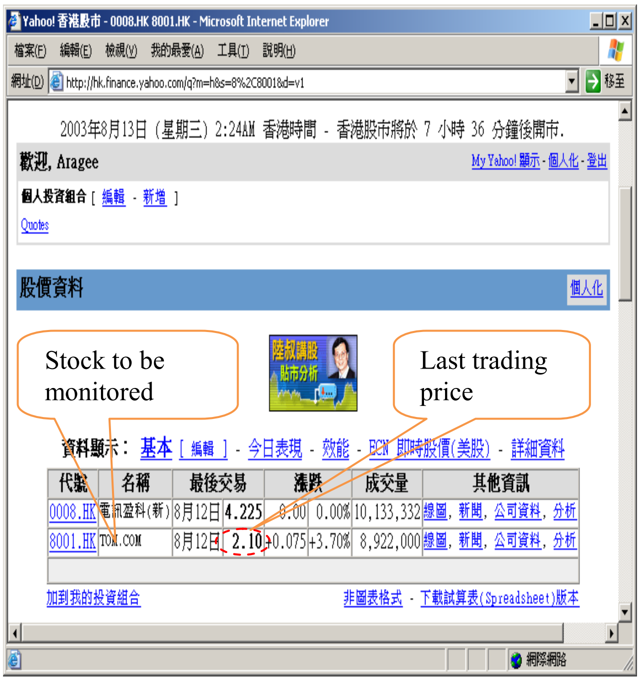
Empirical results : Empirical results obtained show that in the first series of experiments, a *WBA* adopting *WBA4* (i.e., scanning a Webpage for exact words, synonyms, hyponyms and hypernyms) achieved the minimum mean square error (*MSE*) relative to human users' rating when rating the relevance of Websites. Whereas space limitations preclude all results from being included here, the results showing the *MSE* between users and the *WBA* in the experiments when the *WBA* searched for related words using exact words, synonyms, hyponyms and hypernyms (i.e., *WBA4*) and when the *WBA* only searched for exact words (i.e., *WBA1*) is shown in Fig. 5. The results showing that *WBA4* attained lower *MSEs* than *WBA1* suggest that a *WBA* is more likely to reduce its error in rating URLs if it is programmed to recognize related words including exact words, synonyms, hyponyms and hypernyms.

For the five combinations of weightings shown in Table II, empirical results in the second series of experiments show that the *WBA* achieved the minimum MSE when it adopted combination 2 in Table II. This generally suggests that a *WBA* is more likely to reduce its error in rating URLs if it is programmed to place almost equal emphasis on all the three heuristics (*OR*, *FO* and *NK*).

VI. Proof-of-Concept Examples

Two examples are provided in this section to illustrate the major functionalities of *WMAs and PWAs*.

Example 1: A *WMA* was deployed to monitor the changing value of the data value associated with "stock.com" in a Webpage shown in Fig. 6.



http://hk.finance.yahoo.com/q?m=h&s=8%2C8001&d=v1

2003年8月13日 (星期三) 2:24AM 香港時間 - 香港股市將於 7 小時 36 分鐘後開市.

歡迎, Aragee   My Yahoo! 顯示 - 個人化 - 登出

個人投資組合 [編輯 - 新增]

Quotes

股價資料   個人化

資料顯示： 基本 [編輯] - 今日表現 - 效能 - ECN 即時股價(美股) - 詳細資料

代號	名稱	最後交易		漲跌		成交量	其他資訊
0008.HK	電訊盈科(新)	8月12日	4.225	0.00	0.00%	10,133,332	線圖, 新聞, 公司資料, 分析
8001.HK	TOM.COM	8月12日	2.10	+0.075	+3.70%	8,922,000	線圖, 新聞, 公司資料, 分析

加到我的投資組合   非圖表格式 - 下載試算表(Spreadsheet)版本

Fig. 6. A Website to be monitored

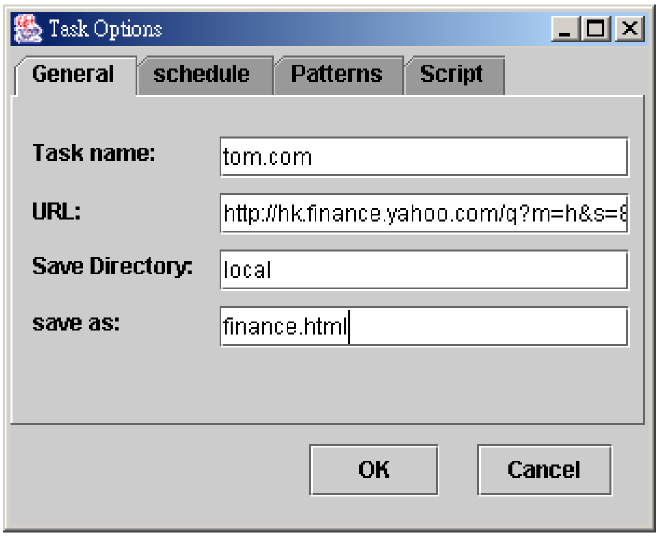
Step 1: In this example, the stock value of tom.com will be monitored, and the user enters the pattern to be monitored (i.e., "tom.com") and the URL http://hk.finance.yahoo.com/q?m=h&s=8001&d=v1 using the interface screen shown in Fig. 7.

Step 2: Subsequently, the user uses the pattern builder interface of the *WMA* to specify the location of the data value associated with "tom.com". In Fig. 6, it can be seen that the data value (i.e., "2.10") of "tom.com" and the pattern "tom.com" is separated by 1) some whitespaces (this is represented in regular expression in Fig. 8 as "\s*", i.e., zero or more whitespace(s)), and 2) a sequence of characters followed by at least one space character (this is represented in regular expression in Fig. 8 as "\S+\s+"). In Fig. 8, "(\S+\s+)" represents the data value to be monitored.

Step 3: The user specifies the monitoring rule using a *WMA*'s script builder shown in Fig. 9. The instruction in Fig. 9 indicates that the *WMA* should notify the user if the stock value of tom.com is above 2.

Step 4: The user specifies the monitoring interval using a *WMA's* schedule interface shown in Fig. 3.

Step 5: When the stock value of tom.com is above 2, the *WMA* notifies the user as shown in Fig. 10.



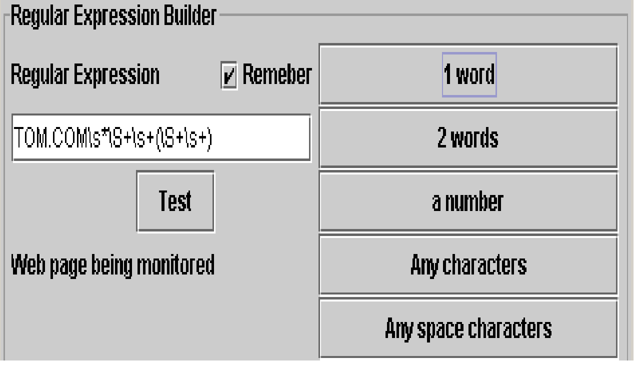
Task name: tom.com

URL: http://hk.finance.yahoo.com/q?m=h&s=8

Save Directory: local

save as: finance.html

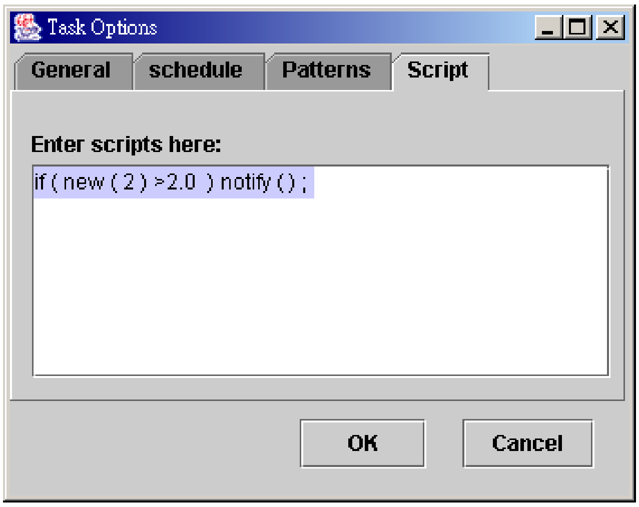
Fig. 7. WMA Task Specification Interface.



Regular Expression ☑ Remeber

TOM.COM\s*\S+\s+(\S+\s+)

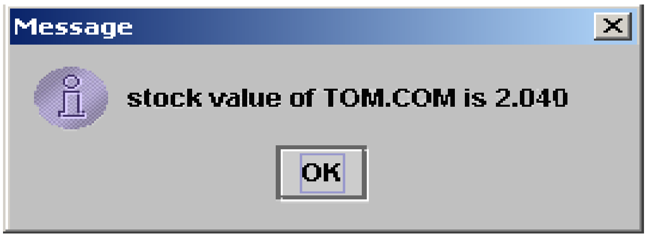
Fig. 8. *WMA* Pattern Builder Interface



Enter scripts here:

```
if ( new ( 2 ) >2.0 ) notify ( ) ;
```

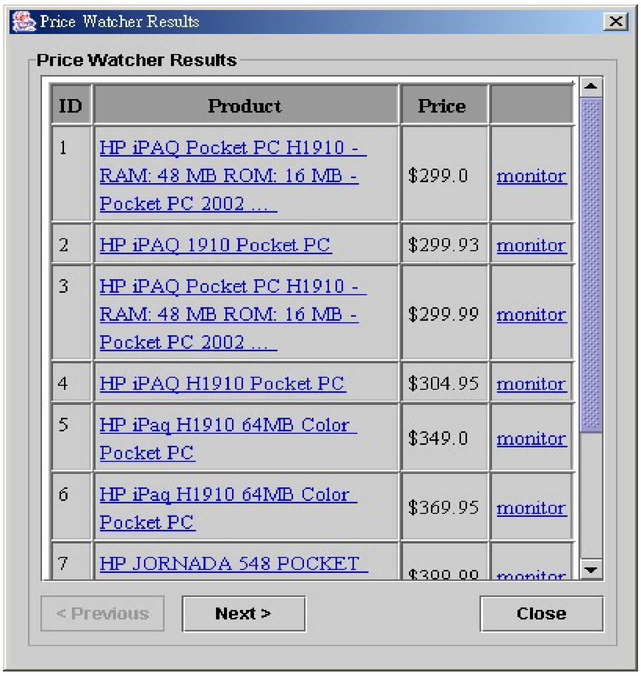
Fig. 9. *WMA* Script Builder Interface



stock value of TOM.COM is 2.040

Fig. 10. *WMA* Notifies User

Example 2: A PWA supports a user by invoking a search engine to search for the price of a HP Pocket PC, deploying *WBAs* to browse and determine the relevance of a list of Websites displaying HP Pocket PC, deploying *WMAs* to monitor changes in the prices of HP Pocket PC in selected Websites then displaying in ascending order the prices of HP Pocket PC from different suppliers' Websites (Fig. 11).



Price Watcher Results

ID	Product	Price	
1	HP iPAQ Pocket PC H1910 - RAM: 48 MB ROM: 16 MB - Pocket PC 2002 ...	$299.0	monitor
2	HP iPAQ 1910 Pocket PC	$299.93	monitor
3	HP iPAQ Pocket PC H1910 - RAM: 48 MB ROM: 16 MB - Pocket PC 2002 ...	$299.99	monitor
4	HP iPAQ H1910 Pocket PC	$304.95	monitor
5	HP iPaq H1910 64MB Color Pocket PC	$349.0	monitor
6	HP iPaq H1910 64MB Color Pocket PC	$369.95	monitor
7	HP JORNADA 548 POCKET	$399.99	monitor

Fig. 11. Price Watcher Agent User Interface.

VII. Discussion and Conclusion

This paper has presented an enhanced holistic *IR* system. It serves the emphasis to mention that the system in this project is not designed to replace or compete with existing search engines. Rather, it is designed to augment and complement the functionalities of existing search engines.

The novel features of this project are as follows. Multiple *WBAs* can be deployed in parallel to simultaneously visit, browse, and scan the information contents of multiple Websites and autonomously determine and rate the relevance of the contents in multiple Websites. Multiple *WMAs* can be deployed in parallel to simultaneously monitor, track, and report changes in multiple Websites. Whereas preliminary ideas of holistic *IR* were reported in [4,6], this work extends the work in [4,6] as follows. Whereas the information filtering agents in [4,6] *only* rate the relevance of Webpages by considering exact words, hyponyms, and hypernyms, *WBAs* in this work 1) consider meronyms and holonyms when determining the relevance of Webpages and 2) consider exact words, hyponyms, and hypernyms when scanning and rating the relevance of Webpages. Whereas *only* the general algorithm of Webpage monitoring was presented in [4,6], *WMAs* in this work are built with user interfaces for specifying information monitoring rules, pattern and data value to be monitored and schedule for monitoring. Additionally, *PWAs* were not considered in [4-6]. *PWAs* in this work can be viewed as "meta-software- agents" invoking on the functionalities of *WBAs* and *WMAs* for browsing and monitoring multiple Websites containing prices of a product.

Acknowledgment

Thanks to Chan Siu Man and Cheung Ting Yee for carrying out the implementations.

References

[1] G.A. Miller. WORDNET: An On-line Lexical Database. International Journal of Lexicography 3-4, pages 235-312.

[2] N. Fridman and C. Hafner. The State of the Art in Ontology Design. AI Magazine, Fall 1997, pp. 53-74.

[3] D. Ellis. "A Behavioral Model for Information Retrieval System Design". J. of Documentation, vol. 49, no. 4, pp. 356-369, 1989.

[4] K. M. Sim and P.T. Wong. Towards Agency and Ontology for Web-based Information Retrieval. IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, Vol. 34, No. 3, 2004, pp. 1-13.

[5] K. M. Sim. Toward an Ontology-enhanced Information Filtering Agent. ACM SIGMOD Record, Vol 33, No. 1, March 2004, pp. 95-100.

[6] K. M. Sim. "Towards Holistic Web-based Information Retrieval: An Agent-based Approach". In Proc. of the 2003 IEEE/WIC Int. Conf. Of Web Intelligence, pp. 39-46.

[7] Winston P. and Chaffin R. A Taxonomy of Part-Whole Relations. Cognitive Science, No. 11, pp. 417-44, 1987.

[8] G. Salton. Automatic Text Processing. Addison Wesley, 1989.

[9] D. L. McGuinness. Ontological Issues for Knowledge-Enhanced Search. In Proc. of Formal Ontology in Information Systems, pp. 302-316, 1998.