# A Personal Agent for Chinese Financial News on the Web

**Christopher C. Yang and Alan Chung**
*Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong. E-mail: yang@se.cuhk.edu.hk*

**As the Web has become a major channel of information dissemination, many newspapers expand their services by providing electronic versions of news information on the Web. However, most investors find it difficult to search for the financial information of interest from the huge Web information space–information overloading problem. In this article, we present a personal agent that utilizes user profiles and user relevance feedback to search for the Chinese Web financial news articles on behalf of users. A Chinese indexing component is developed to index the continuously fetched Chinese financial news articles. User profiles capture the basic knowledge of user preferences based on the sources of news articles, the regions of the news reported, categories of industries related, the listed companies, and user-specified keywords. User feedback captures the semantics of the user rated news articles. The search engine ranks the top 20 news articles that users are most interested in and report to the user daily or on demand. Experiments are conducted to measure the performance of the agents based on the inputs from user profiles and user feedback. It shows that simply using the user profiles does not increase the precision of the retrieval. However, user relevance feedback helps to increase the performance of the retrieval as the user interact with the system until it reaches the optimal performance. Combining both user profiles and user relevance feedback produces the best performance.**

## Introduction

As the popularity of World Wide Web increases, most commercial companies have made the Web to be their major channel of information delivery. It has been estimated that the amount of information on the Internet double every 18 months. Traditional newspapers are expanding their services by providing on-line news on the Web. Comparing to the traditional "ink-on-paper" newspapers, the Web provides real time dissemination of news. Readers no longer need to wait until the next day to read the most recent news on the day. To the investors, real-time financial news is particularly important for decision making on their investment. Because information on the Web is updated frequently, information overload becomes a significant problem. Most users find it difficult to search for the information they need, although it is so easy to be accessed. Most commercial search engines take keywords as inputs. However, they suffer in low precision and recall. Users end up wasting a lot of time surfing on the Web but do not get anything meaningful. Besides, users without much experience in text retrieval may also have difficulties in choosing the right keywords for their query. Search engines that are able to learn user preferences and search on behalf of the users without users taking too much effort to make the query are desired.

To develop a high-performance personal agent (search engine) for Web financial news, a good understanding of relevance is needed before we can design the mechanism of the agent. The study of relevance has a long history. Mizzaro (1997) has given a thorough review of the literature. Relevance is commonly accepted as a relation between two entities or two groups. In the first group, there are three entities: (1) document, (2) surrogate, and (3) information. Documents are the physical entity where information retrieval systems usually provide as result. Surrogates are representation of document, making up of title, keywords, authors names, abstract, etc. Information is what users receive when reading a document. In the second group, there are four entities: (1) problem, (2) information need, (3) request, and (4) query. Problems are what users require information to solve. Information needs are representation of the problems in the mind of users. Requests are representation of the information needs in a natural language. Queries are representation of the information needs in a system language. Figure 1 illustrates the relations between two groups of entities.

As illustrated in Figure 1, most commercial Web search engines only determine the relevant documents by the relation between documents and queries, where queries are represented by keywords. Unless the users are familiar with the document and query representations and the searching



DOI: 10.1002/asi.10020

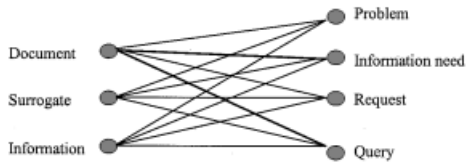


FIG. 1. Releance: relation between two groups of entities.

mechanism of the information retrieval system, it is very difficult to make an appropriate query; especially, most of the Web users are not familiar with computing technologies. A personal agent that is able to represent the user information need without user's effort to make the query to represent his/her own interest is desired. The search engine will then search on behalf of the users to determine the relevant documents based on the relation between information needs and documents. The agent is a software program that operates autonomously and accomplish unique task without direct human supervision. Similar to our human counterparts such as real estate agents and travel agents, the human agents search for the house or traveling plan as the customers give them their preference and interest. The customers may not fully understand the representation of the products; however, the agents learn the user interest by the customer profile and continuous interaction with the customers.

## Literature Review

The Internet search engines act as "spiders" on the Web to search for relevant information (Etzioni & Weld, 1994). Two major approach of Web spiders are: (a) on-line database indexing and searching, and (b) client-based searching agents.

### *On-Line Database Indexing and Searching*

On-line database indexing and searching is the traditional approach. It is based on the database concept of indexing and keyword searching. Systems using this approach collect complete or partial Web documents and then index these documents by keywords on the host server. Searchable interfaces are provided for users to submit their queries. For examples, Lycos, Alta Vista, and Yahoo are using this approach.

Lycos (Mauldin & Leavitt, 1997), developed at Carnegie Mellon University, uses a combination of spider fetching and simple owner registration. Lycos adopts a heuristic-based indexing approach based on title, headings, subheadings, 100 most important words, first 20 lines, size in bytes, and number of words. Alta Vista, developed at Digital's Research Laboratories, provides a full-text index. Alta Vista's success is mainly due to its superior hardware platforms and high-end communication bandwidth. Yahoo partitions the Web into meaningful subject categories. However, the manually created subject categories are cumbersome, time-consuming, and limited in granularity. The demand to create up-to-date and fine-grained subject categories has significantly hindered Yahoo's success. In general, most of these search engines based on the on-line database indexing and searching suffer in poor performance because the Web is growing exponentially and the Web pages are updated frequently. The search engines can hardly cover the huge information space of the Web and update the database indexing as frequent as the Web pages are updated.

### *Client-Based Searching Agents*

Most recent research in Web searching focuses on the development of client-based intelligent searching agents to search for relevant Web pages. The development is either advance artificial intelligence techniques for searching or learning users preferences to enhance searching performance.

*Searching techniques.* Many traditional artificial intelligence techniques have been applied in Web searching. TueMosaic (developed at the Eindhoven University of Technology, TUE) (DeBrat Post, 1994), WebCrawler (purchased by American Online in 1995) (Pinkerton, 1994), and RBSE (Repository Based Software Engineering, funded by NASA) spider, investigate different conventional best first search, depth-first search or breadth-first search. Users submit keywords, specify the depth and width of search for links contained in the current displaying Web pages, and request the spider to fetch the connected Web pages. The relevance of a link is determined based on the similarity between the user's query and the anchor text or full text. Smart Itsy Bitsy Spider (Chen, Chung, Ramsey, & Yang, 1998; Yang, Yen, & Chen, 2000) employs the genetic algorithm and hybrid simulated annealing for searching. WebAnts develops distributed agents to share the indexing loads and searching results to minimize the effort of each agent.

*Learning user preferences.* Other searching agents focus on learning user preferences and recommending Web pages. WebWatcher, Anatagonomy, Syskill, and Webert, Leitizia, and CiteSeer are some prominent examples.

WebWatcher (Armstrong, Freitage, Joaching, & Mitchell, 1995) provides interactive advice to users when they are traversing through the Web links. It incorporates machine-learning methods to acquire knowledge for selecting an appropriate hyperlink on the current visiting Web page. The anchor text, the words in the sentence that containing the hyperlink, the words in the headings, and the words submitted by users are used as the knowledge about the Web page. Full text of documents is not used for retrieval.

Anatagonomy (Kamba, Sakagami, & Koseki, 1997) applies both explicit feedback and implicit feedback to learn user preferences for WWW-based newspaper articles. User scores on each article are used as explicit feedback while the scrolling and enlarging operations are used as implicit feed-

back. The scoring engine rates the articles by comparing the document vector and the user profile. However, users are required to register a set of keywords for each article explicitly and the implicit feedback by scrolling and enlarging operations are not directly corresponding to user interests.

Syskill and Webert (Pazzani & Billsus, 1997; Pazzani, Muramatsr, & Billsus, 1997) applies a naive Bayesian classifier for learning and revising user profiles to determine the interesting Web sites on a given topic. The supervised learning algorithms require a set of positive examples and a set of negative examples. These examples are Web pages that one is interested or not interested in.

Leitizia (Lieberman, 1995, 1997) browses concurrently with users, searches and analyzes Web pages while users are browsing, and displays recommendations continually. A breadth-first search rooted from user's current position is concurrently searching for Web pages.

CiteSeer (Giles, Bollacker, & Lawrence, 1998) indexes academic literature in electronic format, which are usually Postscript files on the Web. CiteSeer autonomously locates, parses, and indexes articles found on the Web. It indexes preprints and technical reports as well as journal and conference papers.

Although many searching agents are developed, they either still require extensive users' effort in revising the queries or do not utilize the user profile and user feedback effectively. The user profile may not capture the necessary information of individual user's interest particularly for their goals of searching. User feedbacks on the rated Web pages are not fully utilized.

### *Intelligent Personal Searching Agents Based on User Profiles and User Feedback*

In this article, we present a personal agent for searching Chinese financial news articles on the Web. User profiles are designed to capture the basic knowledge on user preferences, areas of interest, and reading habits. User feedback is utilized to capture more specific user preferences based on the semantics of the rated news articles. The search engine will then search for the financial news articles that users are most interested in based on the user profiles, user feedback, and the indexed news articles. A Chinese indexer is developed specially for the unknown words in the news articles.

## User Profiles and User Feedback

Compared to the traditional database indexing and searching approach, our system requires less effort from users to specify their query. It learns the user information needs and user preferences from their profiles and their feedback of the rated articles. The user profile is designed for the Chinese financial news to fit the investor's information needs of their investment portfolio. Degree of relevance is considered in the user relevance feedback. The performance of searching increases gradually from time to time until it reaches an optimal performance.

### *User Profiles*

User profiles can be categorized into personal profiles and community profiles (Shepherd, Duffy, Watters, & Gugle, 2001). Personal profiles and community profiles may be used depending on the purpose of news reading. There are two behavioral theories for news reading: (1) *uses and gratification,* and (2) *play or ludenic.* The uses and gratification theory assumes that readers have some underlying goal outside the reading itself. Personal profiles are more appropriate is such situation. The play or ludenic theory, introduced by Stephenson (1967), states that news reading is intrinsically pleasurable, which is more casual, spontaneous, and unstructured (Dozier & Rose, 1984). Such kind of recreational news reading relies mainly on browsing for information seeking because there is no specific goal to achieve. In this case, a personal profile may not be helpful. A community profile is more applicable (Shepherd et al., 2001). Our Chinese Web financial retrieval system retrieves the Chinese news articles that are most recently disseminated on the newspaper Web sites. Users who use the system have a specific goal to retrieve the news articles that contains relevant information for their investment portfolios.

User profile is a structured representation of the user's information needs. Amato and Straccia (1999) classify user profile into five categories: (1) personal data category, (2) gathering data category, (3) delivery data category, (4) actions data category, and (5) security data category. The personal data category is a collection user's personal identification data. The gathering data category includes three subcategories—document content category, document structure category, and document source category. The delivering data category includes two subcategories—delivering means category, and delivery time category. The actions data category involves repeated interactions with users and relevance feedback. The security data category establishes the conditions under which data represented in the user profile may be accessed. However, most of the existing systems represent user profiles by a set of feature of vectors where each element is a keyword. For example, Pazzani and Billsus develop their Syskill and Webert's user profile by selecting a set of informative words using an information-based approach (Pazzani et al., 1997). In our system, we focus on building a user profile that captures user preference on Chinese Financial information published in Hong Kong. Therefore, we construct user profiles by (a) sources of news articles (*SOURCES*), (b) regions of news (*REGIONS*), (c) categories of industries (*INDUSTRIES*), (d) listed companies in HK stock market (*COMPANIES*), and (e) user-specified keywords (*KEYWORDS*). The categories of industries, listed companies, and user-specified keywords collect the personal data. The sources of news articles and regions of news collect the gathering data. For the delivery

TABLE 1. Chinese newspaper sources

| Newspaper Source | | URL |
|---|---|---|
| Apple Daily Online | 蘋果日報 | http://www.appledaily.com.hk |
| Ming Pao Electronic News | 明報 | http://www.mingpao.com/newspaper |
| Oriental Daily News | 東方日報 | http://www.orientaldaily.com.hk |
| Hong Kong Commerce Daily | 香港商報 | http://www.hkcd.com.hk |
| Sing Tao Electronic Daily | 星島電子日報 | http://www.singtao.com |
| Ta Kung Pao | 大公報 | http://www.takungpao.com.hk |

data, the system delivers the retrieved news articles to users daily or on demand where a graphical user interface is developed using Java. For the action data, user relevance feedback will be obtained from users when they read the news articles. For the security data, all the data of user profile is saved on the client machine, where only the agents personalized for each individual user is able to access the profile.

### *User Feedback*

The personal user profile captures the initial knowledge of user preference in general; however, the user preference on the specific content obtained from each news article is not captured. In our system, we use the user-relevance feedback to obtain additional information on user preference. Feedback facilitates communication between users and information retrieval systems, which involves user evaluation of the retrieval output, user judgment and query modification (Spink, 1997a). User-relevance feedback provides specific information of users interest obtained from the rated news articles. It has been reported that user-relevance feedback provides large improvement on information retrieval performance (Salton & Buckley, 1990).

There are two major models of feedback: cybernetic models, and social models. Cybernetic models consider feedback as a closed loop of signed circular causality underlying automatic control processes; social models view feedback as a loop of mutual causality underlying fundamental social processes (Spink, 1997b). Spink and Greisdorf (2001) has investigated the regions across a distribution of user's-relevance judgments, including categorizing, measuring, and evaluating these regions. A dichotomous measure (relevant/not relevant) is used traditionally (Taube, 1965); however, Sperber and Wilson (1986) suggest that relevance must be accounted for the degrees of relevance. It is more important why users accept or reject but not simply the outcome of accept or reject (Saracevic, 1975).

In our Chinese Web financial news retrieval system, we adopt the categorical measurement with five categories: not relevant, partially not relevant, partially relevant, relevant, and perfectly relevant.

Details of implementation of user profiles and user feedback are discussed in the User Profiles and User Feedback Sections.

## System Architecture and Design

The system architecture of the intelligent Chinese Web financial news retrieval system consists of five components: fetching, indexing, user profile, feedback, and search engine. The fetching and indexing components fetch the daily financial news articles from the newspaper Web sites and index each fetched document. The user profile captures the knowledge of user preference on the financial news. The user feedback captures the semantics of the user-interested documents obtained by the search engine. The search engine retrieves the relevant documents based on the indexing of the documents, the user profile and the user feedback. Figure 2 illustrates the architecture of the Chinese Web financial news retrieval system.

### *Fetching*

The system monitors the sources of Chinese financial news on the Web and downloads the most recent published news articles. The sources of the financial news that are currently monitored by our system are listed in Table 1. The number of sources is not limited and is easily expanded in our system.

Several fetching programs, such as Lynx and Html-Gobble, are available on the Web to fetch and display

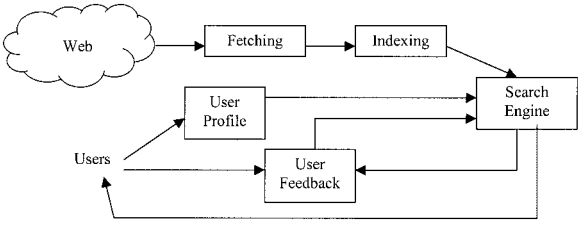


FIG. 2. System architecture of the intelligent Chinese financial news retrieval system.

HTML documents. To make our system more portable and integrate with other components, we implement a generic fetching program in Java. It takes the Universal Resource Location (URL) of the Web page, and uses the Hyper Text Transfer Protocol (HTTP) to make the connection to the corresponding Web site.

*Chinese Indexing*

A traditional indexer recognizes and selects essence of a document and represents it, which is very important in information retrieval. Much research has been done on English indexing, however, there are relatively less on Chinese indexing. The smallest indexing units in Chinese documents are words, while the smallest units in Chinese sentence are characters. Unlike English text, Chinese text has no delimiter to mark word boundaries. In English, spacing often reliably indicates word boundaries. However, in Chinese, a number of characters are placed together without any delimiters, indicating the boundaries between consecutive characters. For example, "international financial center" is translated to "國際金融中心", where "international" is translated to "國際," "financial" is translated to "金融," and "center" is translated to "中心." However, there is not any delimiter between "際" and "金," and "融" and "中" to determine the boundaries between the consecutive characters. The lack of delimiters makes Chinese indexing more difficult than English indexing. There are three major approaches on Chinese indexing: (1) statistical approach, (2) lexical rule-based approach, and (3) hybrid approach based on statistical and lexical information. In newspaper articles, names, places, organizational and historical events, and specialized terms are very common. These words are considered as unknown words that do not appear in dictionaries. Therefore, the lexical rule-based approach, which also known as dictionary-based approach, is not appropriate. In this system, we apply the boundary detection (Yang, Luk, Yung, & Yen, 2000; Yang, Yen, Yung, & Chung, 1998) based on the statistical approach. Experimental results have shown that the boundary detection has over 90% accuracy, and is able to identify most of the unknown words, such names of government officers, names of companies, etc.

Mutual information $I(a, b)$ is the statistical measurement of association between two events, $a$ and $b$. In Chinese segmentation, mutual information, $I(c_i, c_j)$, measures association between two consecutive characters, $c_i$ and $c_j$, in a sentence. Characters that are highly associated are considered to be grouped together to form words.

Equation 1 shows the formulation to calculate the mutual information $I(c_i, c_j)$ for two consecutive characters. The frequencies of characters, $f(c_i)$ and $f(c_j)$, divided by the total number of characters in corpus, $N$, correspond to the probabilities of characters, $c_i$ and $c_j$. The frequency of two consecutive characters, $f(c_i, c_j)$, divided by $N$ correspond to the joint probability of two characters, $c_i$ and $c_j$.

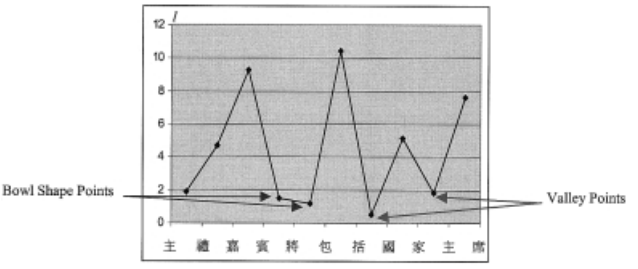


FIG. 3. Mutual information values of the sentence 主禮嘉賓將包括國家主席

.

$$I(c_i, c_j) = \log_2\left(\frac{f(c_i, c_i)/N}{f(c_i)/N\, f(c_j)/N}\right) = \log_2\left(\frac{Nf(c_i, c_i)}{f(c_i)f(c_j)}\right) \quad (1)$$

Mutual information of two characters shows how strongly these characters associated with one another. If the characters are independent to one another, $I(c_i, c_j)$ equals to 0. If $c_i$ and $c_j$ are highly correlated, $I(c_i, c_j)$ increase.

In our boundary detection approach, we detect the boundary of a word by determining if the value mutual information between two characters is lower than a threshold and/or if there is any abrupt change in mutual information.

The *algorithm for boundary detection* is given as:

1. *Counting occurrence frequencies* Obtain occurrence frequencies for all uni-grams and bi-grams.
2. *Compute mutual information for all bi-grams*
3. *Determine the segmentation points:*
   (a) If the mutual information value for a bi-gram is less than a threshold, $T_1$, the point between the two characters in the bi-gram is treated as the segmentation point. $T_1$ is greater than or equal to 0.
   (b) Given a string of characters $\ldots, c_{j-1}, c_j, c_{j+1}, c_{j+2}, c_{j+3}, \ldots,$
4. *Determine the value point:* If $I(c_{i-1}, c_j) > I(c_j, c_{j+1})$ and $I(c_{j+1}, c_{j+2}) > I(c_j, c_{j+1})$ Then the point between $c_j$ and $c_{j+1}$ is a valley point and the point is treated as a segmentation point
5. *Determine the points of bowl shape curve:* If $I(c_j, c_{j+1}) - I(c_{j-1}, c_j) < 0$ and $I(c_{j+2}, c_{j+3}) - I(c_{j+1}, c_{j+2}) > 0$ and $(I(c_{j-1}, c_j) - I(c_j, c_{j+1}))/|I(c_j, c_{j+1}) - I(c_{j+1}, c_{j+2})| > T_2$ and $(I(c_{j+2}, c_{j+3}) - I(c_{j+1}, c_{j+2}))/|I(c_j, c_{j+1}) - I(c_{j+1}, c_{j+2})| > T_2$ where $T_2$ is a threshold Then the points between $c_j$ and $c_{j+1}$ and between $c_{j+1}$ and $c_{j+2}$ are points of bowl shape curve these points are treated as a segmentation point

$T_1$ and $T_2$ are determined experimentally and their values are 1.0 and 2.0, respectively (Yang et al., 1998).

Figure 3 shows an example of the segmentation result. There are two valley points, between 括 and 國 and between 家 and 主. There are two bowl shape points, between 賓 and 將 and between 將 and 包. Taking the valley points and bowl shape points as segmentation

points, the sentence is segmented to five segments, [主禮嘉賓] [將] [包括] [國家] [主席] ([The guest of ceremony] [will] [include] [the country] [chairman]).

Before we apply the boundary detection algorithm to detect word boundaries, we remove the HTML tags of the HTML documents and use the punctuation to segment the document into strings of Chinese characters. The boundary detection algorithm will then be used to segment the of character strings.

After word segmentation, term-weighting heuristics are then computed. Term frequency, $tf_{ij}$, represents the numbers of occurrences of term $j$ in document $i$. The document frequency, $df_j$, represents the number of documents in a collection of $n$ documents in which the term $j$ occurs. The combined weight of term $j$ in a document $i$, $d_{ij}$ is computed as follows:

$$d_{ij} = tf_{ij} \times \log\left(\frac{N}{df_j}\right) \quad (2)$$

The term that occurs more frequent indicates itself as a good descriptor of the document. On the other hand, the term that occurs frequently on many documents implies itself as a general term that does not has any specific meaning. Therefore, a term, which has a high $tf_{ij}$ and low $df_j$, corresponds to a good keyword of the documents.

### *User Profiles*

User profiles capture user preference on Chinese financial information published in Hong Kong newspaper. Five components are included: (1) sources of news articles (*SOURCES*), (2) regions of news (*REGIONS*), (3) categories of industries (*INDUSTRIES*), (4) listed companies in HK stock market (*COMPANIES*), and (5) user-specified keywords (*KEYWORDS*).

PersonalUserProfile =

(SOURCES × REGIONS × INDUSTRIES × COMPANIES × KEYWORDS)

*Sources of news articles* ($w_s$). The system currently uses six newspaper sources on the Internet (Table 1). Different users have different preferences on the information providers. Although similar content are reported by different information providers, investors find some of the authors in some particulars newspapers more reliable and these authors' comments are more helpful in their decision making. Therefore, these investors prefer to read articles form particular newspaper Web site in certain financial issues. A slider on the graphical user interface is provided for users to submit their confidence level, $w_s$, ranged from excellent to very bad for each newspaper source.

$SOURCES = (S_1, S_2, S_3, S_4, S_5, S_6)$
$S_1 =$ *http://www.appledaily.com.hk/*
$S_2 =$ *http://www.mingpao.com/newspaper/*
$S_3 =$ *http://www.orientaldaily.com.hk/*
$S_4 =$ *http://www.hkcd.com.hk*
$S_5 =$ *http://www.singtao.com*
$S_6 =$ *http://www.takungpao.com.hk*
$w_s =$ (*Very Bad*|*Bad*|*Average*|*Good*|*Excellent*)

*Preference on regions of news* ($w_r$). Because Hong Kong is an international financial center, besides local financial news, news from China and international (such as, south east Asia, Pacific region, North America, and Europe) will affect the Hong Kong stock market. In most of the newspaper sources, the financial news is categorized into three regional categories: (1) local, (2) China, and (3) international. For different users, news from different regions may affect their investment by different degree. The user profile of our system captures the importance of the news, $w_r$, from different regions for each user by the slider on the user interface.

$REGIONS = (R_1, R_2, R_3)$
$R_1 =$ *Local Financial News*
$R_2 =$ *China Financial News*
$R_3 =$ *International Financial News*
$w_r =$ (*Very unimportant*|*unimportant*|*Average*|*Important*|*Very Important*)

*Categories of industries.* There are several major industries in Hong Kong. In our systems, we select 10 industries to focus on: (a) real estates, (b) finance, (c) banking, (d) tourism, (e) manufacturing, (f) technology, (g) food and beverage, (h) services, (i) entertainment, and (j) insurance. For each industry, we select a list of keywords (shown in Table 2) that are most significant in the corresponding industry. Users may select the preferred industries by checking the appropriate check box in the panel of the user interface.

$INDUSTRIES = (I_1, I_2, I_3, I_4, I_5, I_6, I_7, I_8, I_9, I_{10})$
$I_1 =$ *Real Estate,* $I_2 =$ *Finance,* $I_3 =$ *Banking,* $I_4 =$ *Tourism,* $I_5 =$ *manufacturing,*
$I_6 =$ *technology,* $I_7 =$ *food and beverage,* $I_8 =$ *services,* $I_9 =$ *entertainment,* $I_{10} =$ *insurance*
$I_i = (k_{i1}, k_{i2}, \ldots), i = 1, 2, \ldots, 10$
$k_{ij} =$ *jth keyword of ith industry*

*Listed companies in Hong Kong stock market.* In our system, user can configure the agent to monitor news articles, which are particularly related to a listed company. Our agent provides a list of company names and their stock codes in the Hong Kong Stock Exchange for users to select as shown in Figure 4.

TABLE 2. List of predefined keywords in industry items

| Industry | Examples of keywords | |
|---|---|---|
| Real Estate 地產業 | *單位 面積 樓宇 房屋 住宅* ... | flats, area, buildings, housing, residence, ... |
| Finance 金融業 | *恆指 期指 基金 聯交所 股價 股票* ... | HS index, index futures, funds, SEHK, stock price, stocks, ... |
| Banking 銀行業 | *外匯 銀行 利率* ... | foreign exchange, banks, interest rates, ... |
| Tourism 旅遊業 | *遊客 酒店 景點* ... | tourists, hotels, scenic points, ... |
| Manufacturing 製造業 | *生產 成衣 製造* ... | production, textile products, manufacturing, ... |
| Technology 科技業 | *高科技 電訊 科研 軟件* ... | high technologies, telecommunication, scientific research, software, ... |
| Food and Beverage 飲食業 | *酒家 酒樓 飲食* ... | restaurants, food and beverages, ... |
| Service 服務業 | *零售 外貿 轉口* ... | retails, exports, exchange, ... |
| Entertainment 娛樂業 | *唱片 偶像 藝人* ... | CDs, idols, actors, ... |
| Insurance 保險業 | *保險 人壽保險 保障* ... | insurance, life insurance, protection, ... |

*User specified keywords.* Besides the categories of industries and the listed companies, users may also specify his or her interests by supplying specific keyword. These interest terms can be person names, locations, company names, etc., in any number of Chinese character or English words. The system provides an interface for the user to edit their keyword list in their profile (Fig. 5). Users can change the keywords any time to adjust their change of preferences or any additional interests. The agent will then match the user-specified keywords with the news articles and count their frequency in each news articles. However, if user did not enter any keyword in this list, the agent will disable this function.

To determine the goodness of a news article in terms of the user profile, a formulation as shown in Equation 3 is adopted. The Personal User Profile Score ($Score_{profile}$) is the accumulation of the relative weight scores obtained from preference on sources of newspapers, regions of news and keywords matching score obtained from categories of industries, listed companies and user-specified keywords. The score of categories of industries and the score of listed companies and user-specified keywords are calculated by dividing the frequencies of the corresponding keywords, $f_{ij}$ and $f_{uj}$, by their cardinalities, $C_i$ and $C_u$, and multiplying to their corresponding weights, $w_i$ and $w_u$.

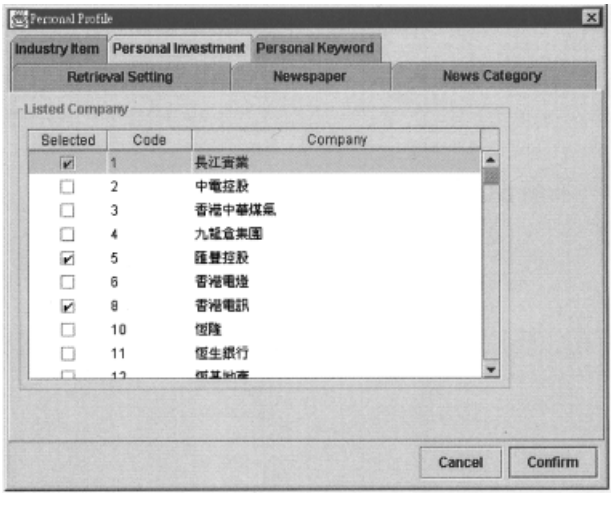


FIG. 4. Listed companies in Hong Kong stock exchange.

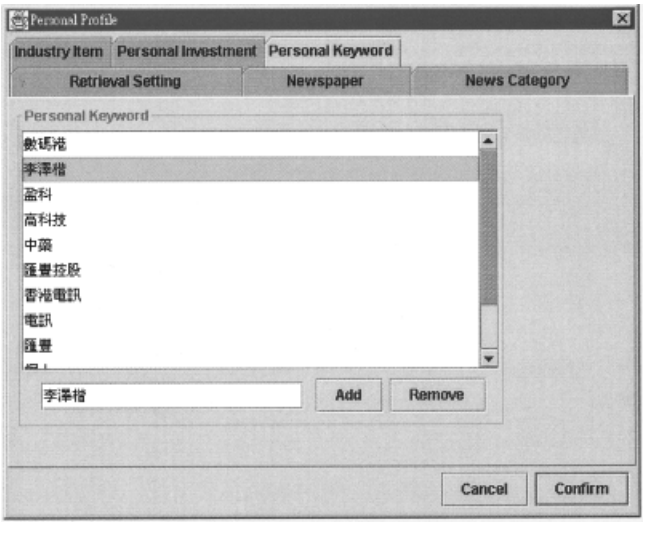


FIG. 5. User-specified keywords.

$$Score_{profile} = w_s \times w_r \times \left( w_i \frac{\sum_j f_{ij}}{C_i} + w_u \frac{\sum_j f_{uj}}{C_u} \right) \quad (3)$$

where $w_s$ is the weight of sources of newspaper; $w_r$ is the weight of regions of news; $w_i$ is the weight of categories of industries; $f_{ij}$ is the frequency of keyword $j$ in categories of industries; $C_i$ is the cardinality of keywords in categories of industries; $w_u$ is the weight of listed companies and user-specified keywords; $f_{uj}$ is the frequency of keyword $j$ in listed companies and user-specified keywords; and $C_u$ is the cardinality of keywords in listed companies and user-specified keywords.

In our implementation, the weights of sources of newspaper and regions of news, $w_s$ and $w_r$, are converted from the five-scale ratings to numerical values of 0.00, 0.25, 0.50, 0.75, and 1.00, respectively. The weight of listed companies and user-specified keywords, $w_u$, is three times the weight of categories of industries, $w_i$, because we find that the user-specified keywords are more specific and important to reflect the user preference of information than the system defined keywords experimentally. $w_u = 3.0$ and $w_i = 1.0$.

### *User Feedback*

User feedback will then be used in the learning mechanism, which is based on the latent semantic structures of the news articles and the past accessed history in terms of the usage of words across documents. Categorical measurement with five categories is adopted to measure the relevance of the retrieved articles.

Relevance = (Not Relevant|Partially Not Relevant|
Partially Relevant|Relevant|Perfectly Relevant)

If two documents are similar in content, the usage of words between these two documents should be similar. Many statistical techniques have been used to estimate this latent structure. In our system, we adopt the Jaccard's similarity function to measure the relevance between the financial news articles that has been rated by user in the previous days and the newly fetched financial news articles. The Jaccard's score between two news articles, $A$ and $B$, is computed as follows:

$$J(A, B) = \frac{\sum_{j=1}^{L} d_{Aj} d_{Bj}}{\sum_{j=1}^{L} d_{Aj}^2 + \sum_{j=1}^{L} d_{Bj}^2 - \sum_{j=1}^{L} d_{Aj} d_{Bj}} \quad (4)$$

where $d_{Aj}$ is the combined weight of term $j$ in article $A$; $d_{Bj}$ is the combined weight of term $j$ in article $B$; and $L$ is the total number of keywords.

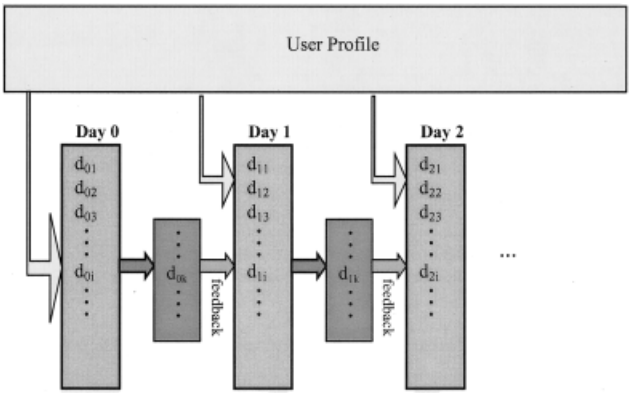


FIG. 6. Retrieval of finance news articles based on user profile and user feedback.

$$d_{ij} = tf_{ij} \times \log \frac{N}{df_j}$$

where $tf_{ij}$ represents the term frequency of term $j$ in article $I$; $df_j$ represents the document frequency of term $j$; and $N$ represents the number of documents accumulating from the first day.

The semantic relevance feedback score, $Score_{feedback}$, is computed as follows:

$$Score_{feedback} = \sum_{i=0}^{n} w_{B_i} \times J(A, B_i) \quad (5)$$

where $w_{Bi}$ is the conversion of the user rating of article $B_i$ to numerical values of 0.00, 0.25, 0.50, 0.75, and 1.00; and $J(A, B_i)$ is the Jaccard's score between the newly fetched article $A$ and the rated article $B_i$; and $n$ is the total number of articles that have been rated on the $m$ previous days, $m = 3$.

### *Search Engine*

As shown on the system architecture in Figure 2, the search engine takes inputs from indexing, user profile, and user feedback. The indexing component determines the keywords for each newly fetched article. The user profile component captures the knowledge of user interest. The user feedback component records the user interest based on the content of each rated article. Based on these inputs, the search engine will rank all the fetched financial news articles on the day and report them to users. However, on Day 0, the search engine has inputs from indexing and user profile only. No articles have been read and rated yet. Starting from Day 1, the search engine will rank all articles based on indexing, user profile, and user feedback, as shown in Figure 6.

For each fetched article, the search engine computes a score based on the user profile score, $Score_{profile}$, and the semantic relevance feedback score, $Score_{feedback}$, as follows:

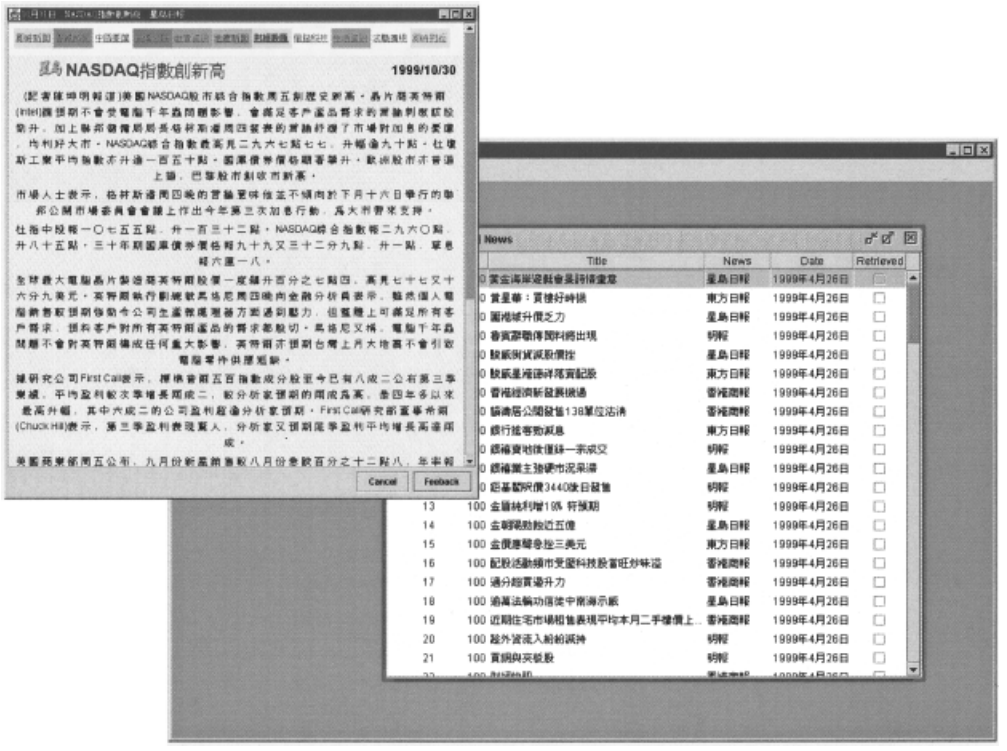


FIG. 7. Result of ranked financial news on a particular day.

$$Score = w_{profile} Score^*_{profile} + w_{feedback} Score^*_{feedback} \quad (6)$$

where $w_{profile}$ is the weighting of user profile score; $w_{feedback}$ is the weighting of semantic relevance feedback score; $Score^*_{profile}$ is the normalized user profile score between 0 and 1; and $Score^*_{feedback}$ is the normalized semantic relevance score between 0 and 1.

The default values of $w_{profile}$ and $w_{feedback}$ are 0.5 so that the impacts of user profiles and user relevance feedback are the same. However, users are able to adjust the weightings. If the weighting of user profile is higher, the learning of user preference from information content will be diminished. On the other hand, if the weight of relevance feedback is higher, the initial performance may not be as good because the preference on information content is accumulated form repeating feedbacks.

The search engine ranks the daily fetched news article based on the score computed by Equation 6. Figure 7 shows the result of the ranked financial news articles on a particular day. When users click on a news title, a news browser will pop up and display the article. A check box next to the news article indicates if the users have read and rated the article. If users prefer not to use such article to be a cue for the retrieval of news article on the next day, they can simply remove the check in the box.

## Experimental Results

We have conducted a user evaluation to examine the performance of the Chinese Web financial news retrieval system based on different setups of user inputs. In the first setup, subjects only provide their user profiles, but the feedback of the ranked articles are not submitted. In the second setup, subjects only provide the ratings of the daily ranked articles, but the initial user profiles are not recorded. In the third setup, subjects provide both user profile and user feedback.

In the user evaluation, 10 subjects from the University of Hong Kong are selected. Each subject is asked to provide his or her user profile and/or feedback and use the system for 5 consecutive days to involve in all three setups. However, subjects did not know the setup they were participating during the experiment. Twenty top-ranked news articles are returned on each day, each subject is asked to determine whether the returned articles are relevant. Approximately, 170 news articles from the six sources of newspapers are fetched everyday. The performance is measured by precision of retrieval.

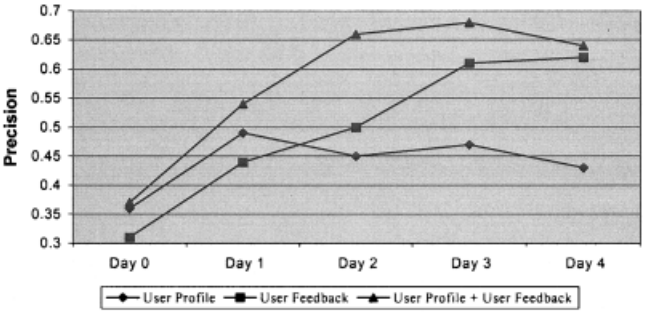


FIG. 8. Experimental results.

$$\text{Precision} = \frac{\text{Number of relevant returned news articles}}{\text{Number of returned new articles}}$$

$$= \frac{\text{Number of relevant returned news articles}}{20} \quad (7)$$

Figure 8 shows the experimental results. For the first setup, for which only user profile is used, the precision increases on Day 1 but does not increase anymore starting from Day 2. It increases slightly on Day 3, but not significantly. In this case, the daily result of the search engine is based on the user profile submitted initially. The input to the search engine does not change from day to day. However, the sets of news articles are different every day. Therefore, the precision only depends on the relevance of the released new articles on the day. For any particular days, if there are relatively fewer articles that are of the interest of users, the precision is comparatively low. The increases or decreases of precision do not reflect the learning ability from day to day. For the second setup, for which only user feedback is used, the precision increases consistently from Day 0 to Day 3 and increases slightly on Day 4. The input to the search engine is the index of the relevant documents obtained from the user feedback, which is different every day. That means the indexing of the earlier released relevant documents help to increase the precision of the searching result. For the third setup, both user profile and user feedback are used, the precision increases consistently from Day 0 to Day 2, increases slightly on Day 3, and decreases slightly on Day 4. When both user profile and user feedback are use, the performance is significantly better than those of using user profile or using feedback only on Day 1 and Day 2, the performance becomes closer to that of using feedback only but is still significantly better than that of using profile only. The decrease in precision on Day 4 may only due to the lower number of relevant documents on the day comparing to other days as we observe the decrease of precision on Day 4 for the first setup.

As shown in the experimental result, the user profile is helpful to retrieve an initial set of news articles that may fit the user interest. However, the user profile does not improve the retrieval performance because it does not obtain further feedback from users. The user-relevance feedback improves the retrieval performance as the system obtains the user-specific preference from the content of the news articles. The degree of relevance obtained from the rated articles helps to filter the irrelevant new articles and capture the semantics of the relevant articles. It is also observed that poor improvement of retrieval performance is obtained if the user always provides negative feedbacks, i.e., the ratings are always not relevant or partially not relevant. It is due to the lack of information to capture the user preference. Merely providing unrelevant articles can filter the irrelevant information but cannot retrieve the relevant information. We also find that the more user-specified keywords submitted to the user profiles of the system, the better the performance is on the first day. However, there is no significant difference after a few days of interaction with the system by user feedback. Combining the user profile and user feedback produce the best performance. Given an initial set of news articles obtained from the user profile, the user-relevance feedback obtains more specific user interest from the relevant articles. The combination of user profile and user feedback helps the system to reach the optimal performance earlier than simply applying the user feedback.

## Conclusion

We have presented a personal agent for retrieving Chinese Web financial news articles. User feedback and user profiles are utilized to learn the user preferences. User profiles capture the knowledge of user preferences based on sources of news articles, regions of news reported, categories of industries related, listed companies in HK stock market, and user-specified keywords. User feedback captures the semantics of the user-rated news articles. The search engine searches for the Web news articles based on the user preferences and indexing on behalf of users. We have conducted an experiment to compare the performance of retrieval based on different setups of user profiles and user feedback. It shows that user profiles do not help in improving the retrieval performances continuously but capture an initial set of news articles that may be of user interest every day. User feedback helps in improving the retrieval performances continuously but the improvement saturates after a certain time. Combining both user profiles and user feedback is significantly better than using either user profiles or user feedback only. Although the developed system focuses on the Chinese financial information, the techniques may also apply on other domains. However, the design of user profiles may need to be modified.

## Acknowledgments

This project was supported by the Direct Research Grant of the Chinese University of Hong Kong, 2050239.

## References

Amato, G., & Straccia, U. (1999). User profile modelling and applications to digital libraries. Proceedings of the 3rd European Conference on Digital Libraries, Paris, France, September 22–25, 1999, pp. 184–197.

Armstrong, R., Freitage, D., Joachims, T., & Mitchell, T. (1995). WebWatcher: A learning apprentice for the World Wide Web. AAAI 1995 Spring Symposium Information Gathering from Heterogeneous, Distributed Environments, Menlo Park, CA.

Chen, H., Chung, Y., Ramsey, M., & Yang, C.C. (1998). A smart itsy bitsy spider for the web. Journal of the American Society for Information Science, 49 (7), 604–618.

DeBra, P., & Post, R. (1994). Information retrieval in the World Wide Web: Making client-based searching feasible. Proceedings of the First International World Wide Web Conference, Geneva, Switzerland.

Dozier, D., & Rice, R. (1984). Rival theories of electronic newsreading. In: R. Rice (Ed.), The new media (pp. 103–128). London: Sage Publications.

Etzioni, O., & Weld, D. (1994). A softbot-based interface to the internet. Communications of the ACM. 37 (7), 72–79.

Giles, C.L., Bollacker, K.D., & Lawrence, S. (1998). CiteSeer: An automatic citation indexing system. Proceedings of the Third ACM Conference on Digital Libraries, New York, pp. 89–98.

Kamba, T., Sakagami, H., & Koseki, Y. (1997). Anatagonomy: A personalized newspaper on the World Wide Web. International Journal of Human-Computer Studies, 46 (6), 789–803.

Lieberman, H. (1995). Letizia: An agent that assists web browsing. Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, August.

Lieberman, H. (1997). Autonomous interface agents. Proceedings of the ACM Conference on Computers and Human Interface, CHI-97, Atlanta, GA, March.

Mauldin, M.L., & Leavitt, J.R.R. (1994). Web-agent related research at the CMT. Proceedings of the ACM Special Interest Group on Networked Information Discovery and Retrieval, August.

Mizzaro, S. (1997). Relevance: The whole history. Journal of the American Society for Information Science, 48 (9), 810–832.

Pazzani, M., Muramatsr, J., & Billsus, D. (1997). Syskil & Webert: Identifying interesting Web sites. Proceedings of the National Conference on Artificial Intelligence, Portland, OR, pp. 54–61.

Pazzani, M., & Billsus, D. (1997). Learning and revising user profiles: The identification of interesting Web sites. Machine Learning, Dorrdrecht, The Netherlands, Kluwer Academic Publishers, (vol. 27, pp. 313–331).

Pinkerton, B. (1994). Finding what people want: Experiences with the WebCrawler. Proceedings of the Second International World Wide Web Conference, Chicago, IL, October 17–20.

Rocchio, J.J. (1971). Relevance feedback in information retrieval. In: G. Salton (Ed.). The SMART retrieval system: experiments in automatic document processing. Englewood Cliffs, NJ: Prentice-Hall.

Sakagami, H., & Kamba, T. (1997). Learning personal preferences on online newspaper articles from user behaviors. Sixth International on World Wide Web Conference, Santa Clara, CA, April 7–11.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 41 (4), 288–297.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science, 26, 321–343.

Shepherd, M., Duffy, J.F., Watters, C., & Gugle, N. (2001). The role of user profiles for news filtering. Journal of the American Society for Information Science and Technology, 52 (2), 149–160.

Sperber, D., & Wilson, D. (1986). Relevance: Communication and cognition. Cambridge, MA: Harvard University Press.

Spink, A. (1997a). Study of interactive feedback during mediated information retrieval. Journal of the American Society for Information Science, 48 (5), 382–394.

Spink, A. (1997b). Information science: A third feedback framework. Journal of the American Society for Information Science, 48 (8), 728–740.

Spink, A., & Greisdorf, H. (2001). Regions and levels: Measuring and mapping users' relevance judgements. Journal of the American Society for Information Science and Technology, 52 (2), 161–173.

Stephenson, W. (1967). The play theory of mass communication. Chicago: The University of Chicago Press.

Taube, M. (1965). A note on the pseudo-mathematics of relevance. American Documentation, 16, 69–72.

Yang, C.C., Luk, J.W.K., Yung, S.K., & Yen, J. (2000). Combination and boundary detection approaches on Chinese indexing. Journal of the American Society for Information Science, 51 (4), 340–351.

Yang, C.C., Yen, J., & Chen, H. (2000). Intelligent internet searching agent based on hybrid simulated annealing. Decision Support Systems, Special Issue on Intelligent Agents and Digital Community, 28 (3), 269–277.

Yang, C.C., Yen, J., Yung, S.K., & Chung, A. (1998). Chinese indexing with mutual information. Proceedings of the First Asia Digital Library Workshop, Hong Kong, August 6–7.