ICIS 2000 Proceedings

International Conference on Information Systems (ICIS)

December 2000

Intelligent Agents for Retrieving Chinese Web Financial News

Christopher Yang
The Chinese University of Hong Kong

Alan Chung
The Chinese University of Hong Kong

INTELLIGENT AGENTS FOR RETRIEVING CHINESE WEB FINANCIAL NEWS

Christopher C. Yang
Alan Chung
The Chinese University of Hong Kong
Hong Kong

Abstract

As the popularity of World Wide Web increases, many newspapers expand their services by providing news information on the Web in order to be competitive and increase benefit. The Web provides real time dissemination of financial news to investors. However, most investors find it difficult to search for the financial information of interest from the huge Web information space. Most of the commercial search engines are not user friendly and do not provide any tailor-made intelligent agents to search for relevant Web documents on behalf of users. Users have to exert a lot of effort to submit an appropriate query to obtain the information they want. Intelligent agents that learn user preferences and monitor the postings of Web information providers are desired. In this paper, we present an intelligent agent that utilizes user profiles and user 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 will rank the top 20 news articles that users are most interested in based on these inputs. Experiments were conducted to measure the performance of the agents based on the inputs from user profile and user feedback.

1. INTRODUCTION

The World Wide Web has become a major channel for information delivery. It has been estimated that the amount of information on the Internet doubles every 18 months. Traditional newspapers are expanding their services by providing on-line news on the Web. Information on the Web is updated frequently. For these reasons, information overload becomes a significant problem. Most users find it difficult to search for the information they need although it is easy to access. Most commercial search engines use keywords as inputs. However, they suffer from low precision and recall. Users end up wasting a lot of time surfing the Web but do not get anything meaningful. Besides, users without much experience in text retrieval may also have difficulty 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 exerting too much effort to make the query are desired.

There are two major approaches for Internet search engines: online database indexing and searching and client-based searching agents.

1.1 Online Database Indexing and Searching

Online database indexing and searching is the traditional approach. 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 use this approach.

Lycos, 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.

1.2 Client-Based Searching Agents

Most recent research in Web searching focuses on developing client-based intelligent searching agents to search for relevant Web pages on behalf of users.

1.2.1 Searching Techniques

Many traditional artificial intelligence techniques have been applied. TueMosaic (DeBra and Post 1994), WebCrawler (purchased by American Online in 1995) (Pinkerton 1994), and RBSE (Repository Based Software Engineering) spider investigate different conventional best first search. Smart Itsy Bitsy Spider (Chen et al. 1998; Yang et al. 2000b) 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.

1.2.2 Learning User Preferences

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

WebWatcher (Armstrong et al. 1995) provides interactive advice to users when they are traversing the Web links. It incorporates machine learning methods to acquire knowledge for selecting an appropriate hyperlink on the Web page currently being visited. The anchor text, the words in the sentence 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 et al. 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 feedback. The scoring engine rates the articles by comparing the document vector and the use profile. However, users are required to register a set of keywords for each article explicitly and the implicit feedback by scrolling and enlarging operations does not directly correspond to user interests.

Syskill & Webert (Pazzani and Billsus 1997; Pazzani et al. 1997) applies a naive Bayesian classifier for learning and revising user profiles to determine 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 in which one is interested or not interested.

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 et al. 1998) indexes academic literature in electronic format, which is 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.

1.3 Intelligent Searching Agents Based on User Profiles and User Feedback

In this paper, we present an intelligent 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.

Compared to the traditional database indexing and searching approach, our system requires less effort from users to specify their query. It learns user preferences from their profiles and their daily feedback of the rated articles. The accuracy of the searching results increases gradually over time.

2. SYSTEM ARCHITECTURE

The system architecture of the intelligent Chinese 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 site and index each fetched document. The user profile captures the knowledge of user preferences on the financial news. The user feedback captures the semantics of the documents of interest to the user 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.

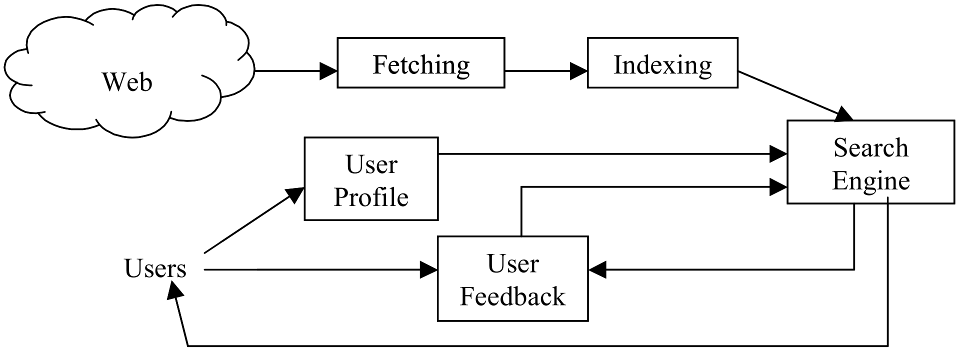
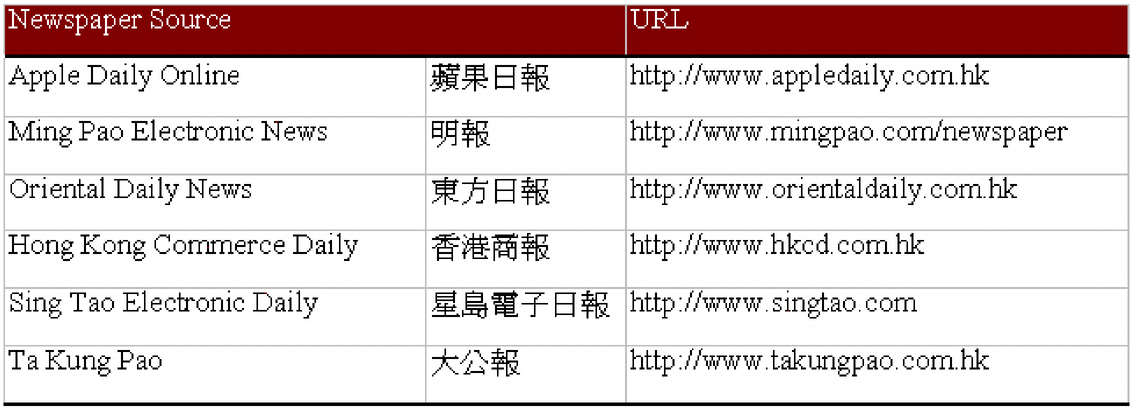


Figure 1. System Architecture of the Intelligent Chinese Financial News Retrieval System

2.1 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.

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

Several fetching programs, such as Lynx and HtmlGobble, are available on the Web to fetch and display HTML documents. In order to make our system more portable and to integrate it 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.

2.2 Chinese Indexing

A traditional indexer recognizes and selects the essence of a document and represents it, which is very important in information retrieval. Much research has been done on English indexing; however, there has been 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. This makes Chinese indexing more difficult than English indexing. There are three major approaches to Chinese indexing, (1) statistical approach, (2) lexical rule-based approach, and (3) hybrid approach based on statistical and lexical information. In this system, we apply the boundary detection (Yang et al. 1998, 2000a) based on the statistical approach.

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.

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

Mutual information of two characters shows how strongly these characters associated with one another. If the characters are independent of one another, $I(c_i,c_j)$ equals 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 of 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 ... , c_{j-1}, c_j, c_{j+1}, c_{j+2}, c_{j+3}, ...,
      Determine the valley point:
         If $I(c_{j-1},c_j) > I(c_j,c_{j+1})$ 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
      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 a bowl shape curve; these points are treated as a segmentation point

Figure 2 shows an example of the segmentation result.

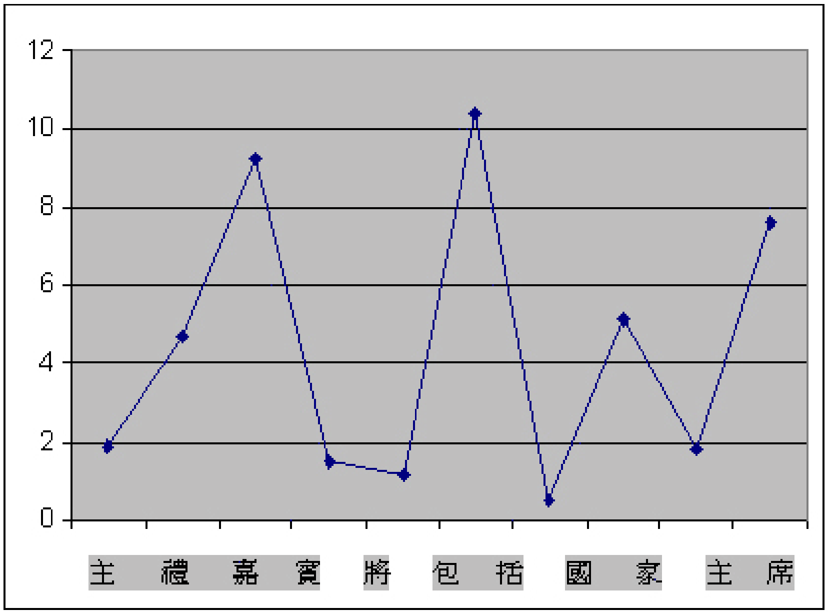


Figure 2. Mutual information values of the sentence 主禮嘉賓將包括國家主席

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 strings of characters.

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:

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

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 have any specific meaning. Therefore, a term, which has a high tf_{ij} and low df_j, corresponds to a good keyword of the documents.

2.3 User Profile

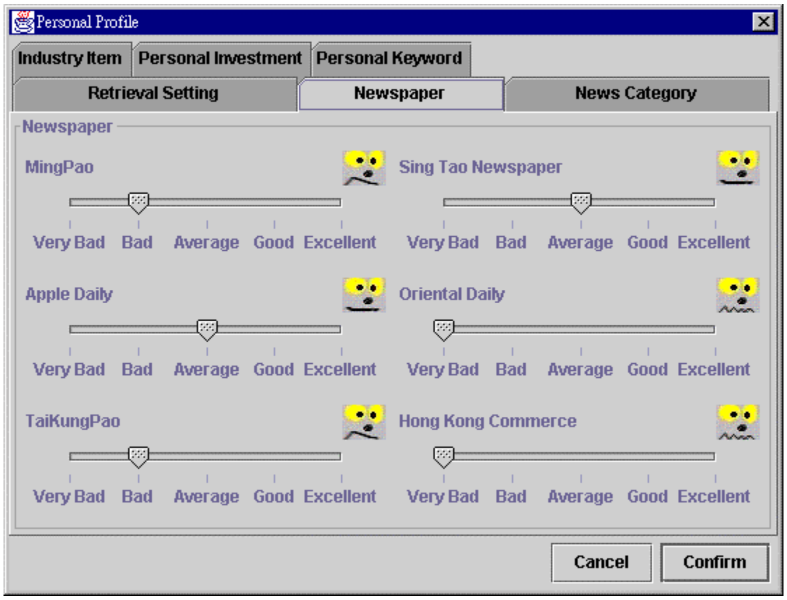
Given the information needs of users and a good information retrieval system, the result of retrieval may still be poor if the user does not provide a query that represents the information needs. A typical query is represented by keywords. Keyword searching obtains high precision and recall only if the user is experienced and knows the right keyword to be used. If the keyword is too general, the retrieval system may return many documents where only a few of them are relevant. That means the precision is low. If the keyword is too specific, the precision is high; however, some other relevant documents that do not use the exact keyword will not be returned as a result. That means the recall is low. Query represented by keywords is rather passive. It requires users to properly present their information needs.

Most users of Web search engines are not experienced users. Many of them are not even experienced users of computers. Financial investors may be knowledgeable in using the financial information to make decisions on their investments, but they may not be proficient in using Web search engines.

In order to obtain the information needs of less experienced users, agents are utilized to build their profiles. The user profile captures basic knowledge on user preferences, areas of interest and reading habits. A good user profile not only increases the precision of retrieval but also narrows the retrieval scope and directly reduces processing time. Our system builds an initial profile by asking users to answer a few questions and explicitly state their preferences for filtering.

Most of the existing systems represent user profiles by a set of feature vectors where each element is a keyword. For example, Pazzani and Billsus develop their Syskill & Webert's user profile by selecting a set of informative words using an information-based approach. 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 (1) sources of news articles, (2) regions of news, (3) categories of industries, (4) listed companies in HK stock market, and (5) user specified 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 is reported by different information providers, investors find some of the authors in particular newspapers to be more reliable and these authors' comments are more helpful in their decision making. Therefore, these investors prefer to read articles from a particular newspaper Web site for certain financial issues. As shown in Figure 3, users may use a slider to submit their confidence level ranged from excellent to very bad for each newspaper source.



Newspaper

MingPao: Very Bad [✓ Bad] Average Good Excellent

Sing Tao Newspaper: Very Bad Bad [✓ Average] Good Excellent

Apple Daily: Very Bad Bad [✓ Average] Good Excellent

Oriental Daily: [✓ Very Bad] Bad Average Good Excellent

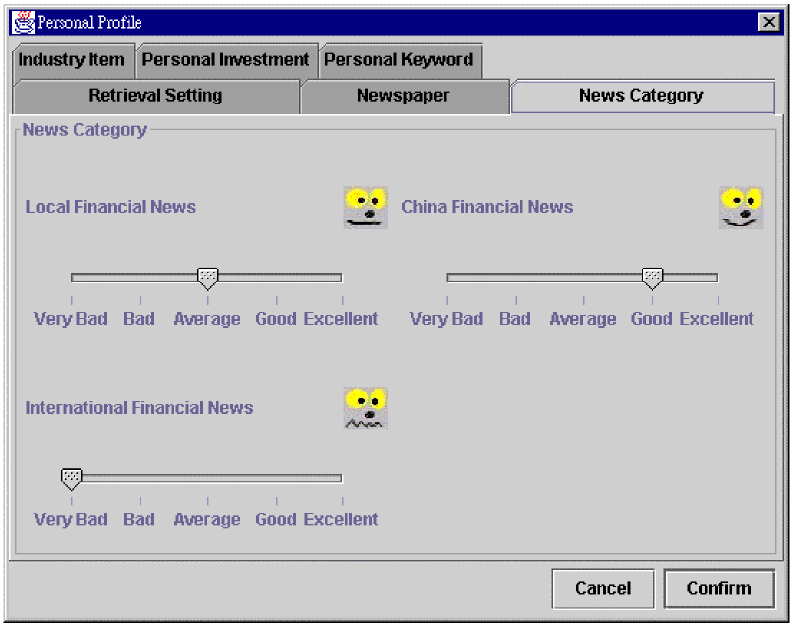
TaiKungPao: Very Bad [✓ Bad] Average Good Excellent

Hong Kong Commerce: [✓ Very Bad] Bad Average Good Excellent

Figure 3. Preferences on Sources of News Articles

Preference on Regions of News (w_r): Since 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 also 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 from different regions for each user by the user interface as shown in Figure 4.

Categories of Industries (f_i): There are several major industries in Hong Kong. In our system, we select 10 industries on which to focus: (1) finance, (2) banking, (3) real estate, (4) technology, (5) manufacturing, (6) services, (7) tourism, (8) entertainment, (9) food and beverage, and (10) 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 as shown in Figure 5.



News Category

Local Financial News

Very Bad Bad Average Good Excellent

China Financial News

Very Bad Bad Average Good Excellent

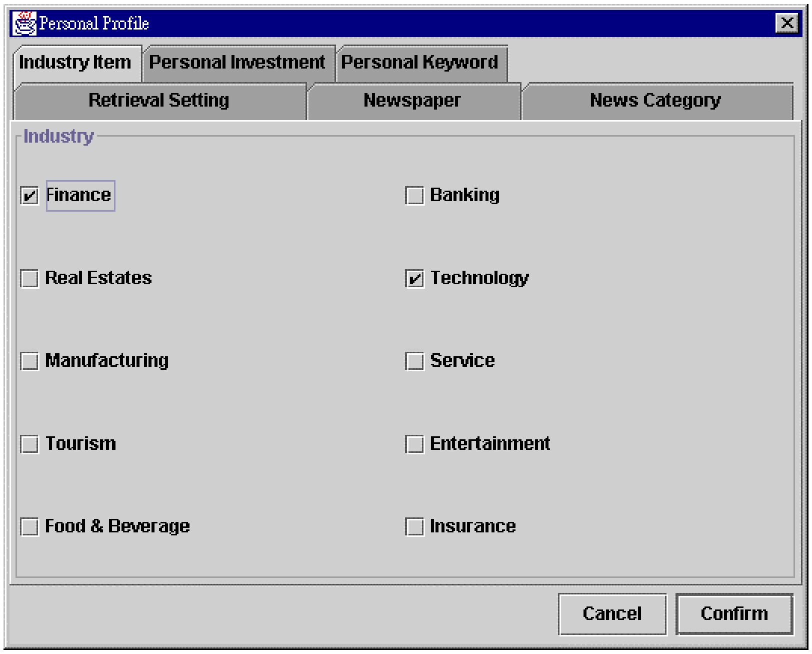
International Financial News

Very Bad Bad Average Good Excellent

Figure 4. Preference on Regions of News

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*
Finance 金融業	*恆指 期指 基金 聯交所 股價 股票*
Banking 銀行業	*外匯 銀行 利率*
Tourism 旅遊業	*遊客 酒店 景點*
Manufacturing 製造業	*生產 成衣 製造*
Technology 科技業	*高科技 電訊 科研 軟件*
Food & Beverage 飲食業	*酒家 酒樓 飲食*
Service 服務業	*零售 外貿 轉口*
Entertainment 娛樂業	*唱片 偶像 藝人*
Insurance 保險業	*保險 人壽保險 保障*

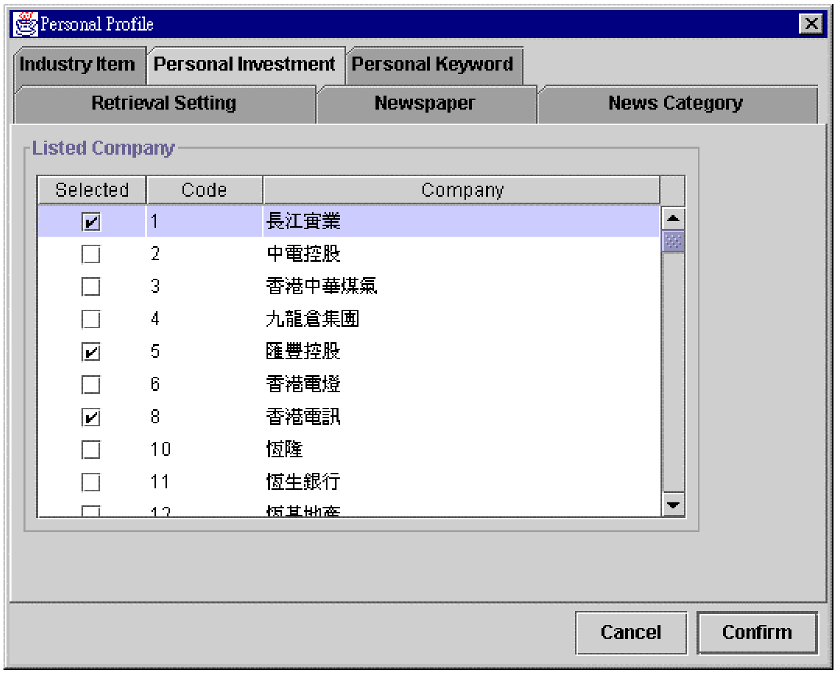


Industry

☑ Finance
☐ Banking
☐ Real Estates
☑ Technology
☐ Manufacturing
☐ Service
☐ Tourism
☐ Entertainment
☐ Food & Beverage
☐ Insurance

Figure 5. Categories of Industries

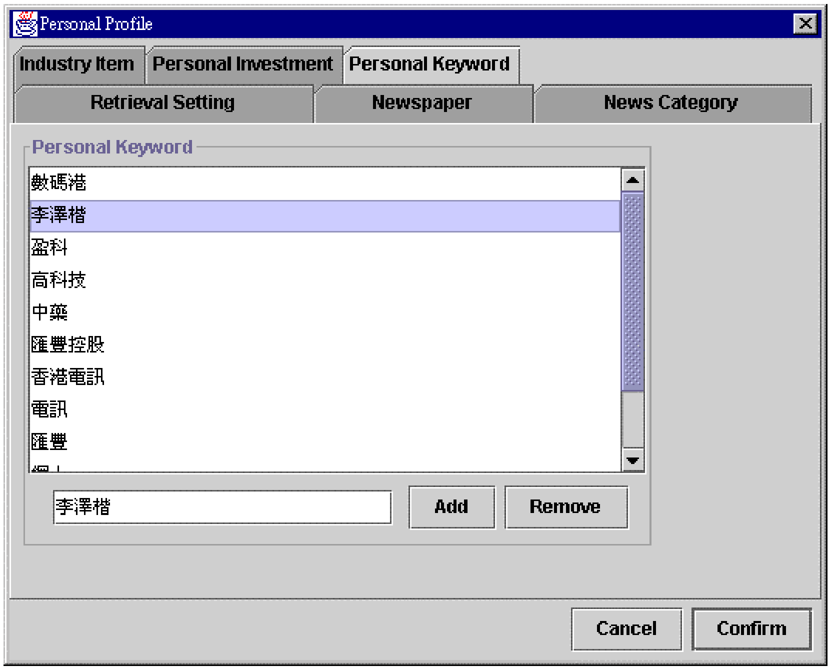
Listed companies in Hong Kong stock market: In our system, user can configure the agent to monitor news articles that are particularly related to a listed company. Our agent provides a list of company names and their stock codes in the Hong Kong Stock Exchange for users to select as shown in Figure 6.



Listed Company

Selected	Code	Company
☑	1	長江實業
☐	2	中電控股
☐	3	香港中華煤氣
☐	4	九龍倉集團
☑	5	匯豐控股
☐	6	香港電燈
☑	8	香港電訊
☐	10	恆隆
☐	11	恆生銀行
☐	12	恆基地產

Figure 6. Listed Companies in Hong Kong Stock Exchange



Personal Keyword

數碼港
李澤楷
盈科
高科技
中藥
匯豐控股
香港電訊
電訊
匯豐

李澤楷

Figure 7. User Specified Keywords

User specified keywords (f_u): Besides the categories of industries and the listed companies, users may also specify their interests by supplying specific keywords. These interest terms can be person names, locations, or company names, etc. in any number of Chinese characters or English words. The system provides an interface for the user to edit their keyword list in their profile (Figure 7). The agent will then match the user-specified keywords with the news articles and count their frequency in each news article. However, if the user did not enter any keywords in this list, the agent will disable this function.

In order to determine the goodness of a news article in terms of the user profile, a formulation (S_p) as shown in Equation (3) is adopted. The User Profile Score (S_p) is the accumulation of the relative weight scores obtained from preference on sources of newspapers, regions of news, and keyword matching scores 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_i and f_u, by their cardinalities, C_i and C_u, and multiplying to their corresponding weights, w_i and w_u.

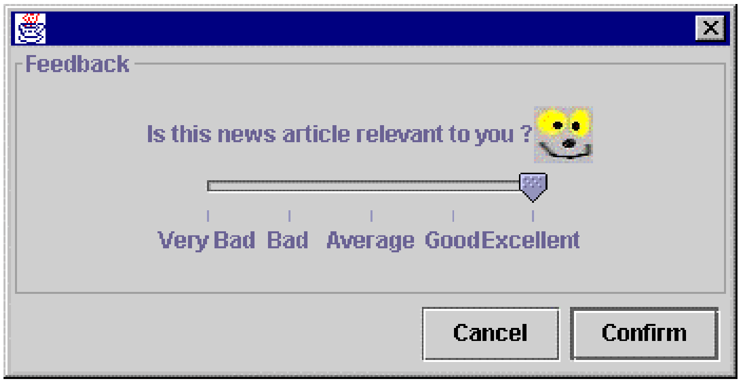
$$S_p = w_s \times w_r \times \left(w_i \frac{\sum_j f_{ij}}{C_i} + w_u \frac{\sum_j f_{uj}}{C_u} \right) \tag{3}$$

where w_s is the weight of the sources of newspaper
w_r is the weight of the 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
C_u is the cardinality of keywords in listed companies and user specified keywords

2.4 User Feedback

The 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 feedback to obtain additional information on user preference. Such feedback provides more specific information about user interest in news topics, events, names, and other relevant knowledge. It has been reported that user relevance feedback provides a large improvement on information retrieval performance (Salton and Buckley 1990).

After reading the ranked articles selected by our system, users may provide feedback to our agent by rating the relevance of the articles. The interface for such feedback is shown in Figure 8. The 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.



Feedback

Is this news article relevant to you ?

Figure 8. User Relevance Feedback Window

We measure the relevance of a newly fetched news article based on the latent semantic structure in terms of the usage of words across documents. 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 than have been rated by the 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:

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

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
L is the total number of keywords

The accuracy of the ranking of newly fetched articles by our agent also relies on the accuracy of user rating on each rated article in user feedback. In other words, if the user provides a high rating for an article in user relevance feedback, the user finds this article interesting and would like to receive more news articles with similar content. Therefore, the semantic relevance score, S_s, is computed as follows:

$$S_s = \sum_{i=0}^{n} w_{B_i} \times J(A,B_i) \qquad (5)$$

where w_{Bi} is the rating of article B_i by user
$J(A,B_i)$ is the Jaccard's score between the newly fetched article A and the rated article B_i
n is the total number of articles that have been rated

2.5 Search Engine

As shown on the system architecture in Figure 1, the search engine takes inputs from indexing, the 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 of the fetched financial news articles on that day and report them to users.

However, on Day 0, the search engine has inputs from indexing and the 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, the user profile, and user feedback.

For each fetched article, the search engine computes a score based on the user profile score, S_p, and the semantic relevance score, S_s, as follows:

$$S = w_p \mathrm{S_p} + w_s \mathrm{S_s} \tag{6}$$

where w_p is the weighting of user profile score
w_s is the weighting of semantic relevance score
$\mathrm{S_p}$ is the normalized user profile score
$\mathrm{S_s}$ is the normalized semantic relevance score

The default values of w_p and w_s are 0.5; however, users are allowed to set their values by the interface shown on Figure 9.

The search engine ranks the daily fetched news articles based on the score computed by Equation (6). Figure 10 shows the result of the ranked financial news articles on a particular day. When the user clicks on a news title, a news browser will pop up and display the article (Figure 11). A check box next to the news article indicates if the user has read and rated the article. If the user prefers that a particular article not be a cue for the retrieval of news article on the next day, they can simply remove the check in the box.

3. EXPERIMENTAL RESULTS

We have conducted a user evaluation to examine the performance of the intelligent Chinese 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 is 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.

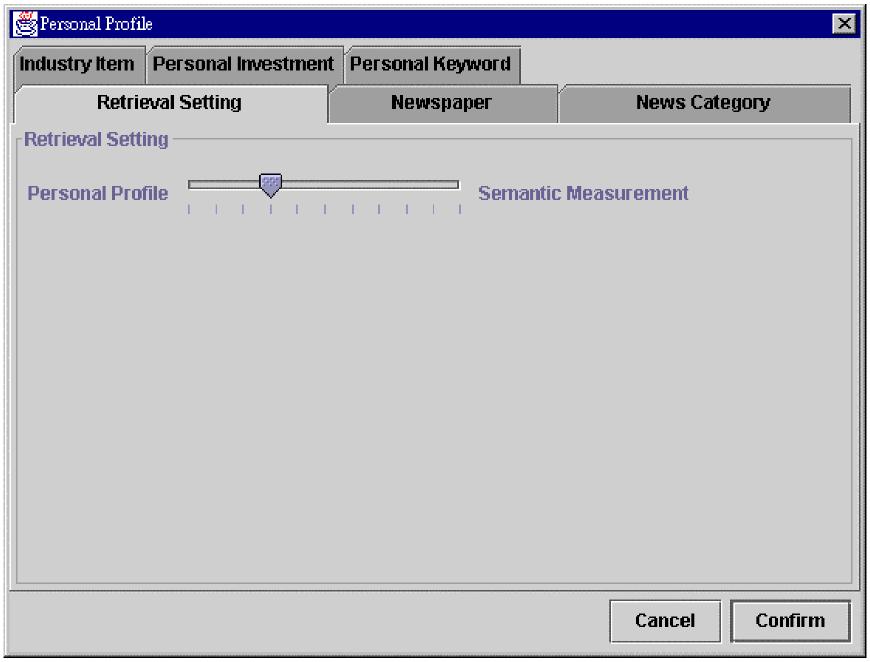
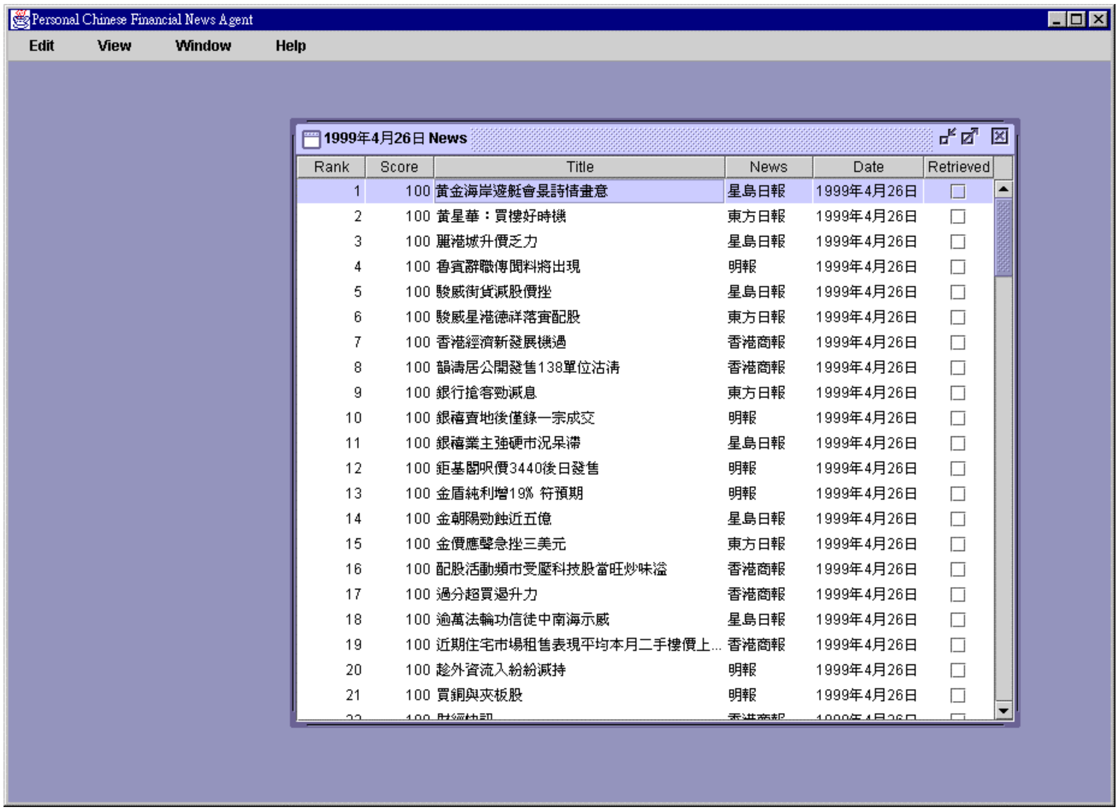


Figure 9. Window for Adjusting the Weightings on User Profile and Semantic Measurement



1999年4月26日 News

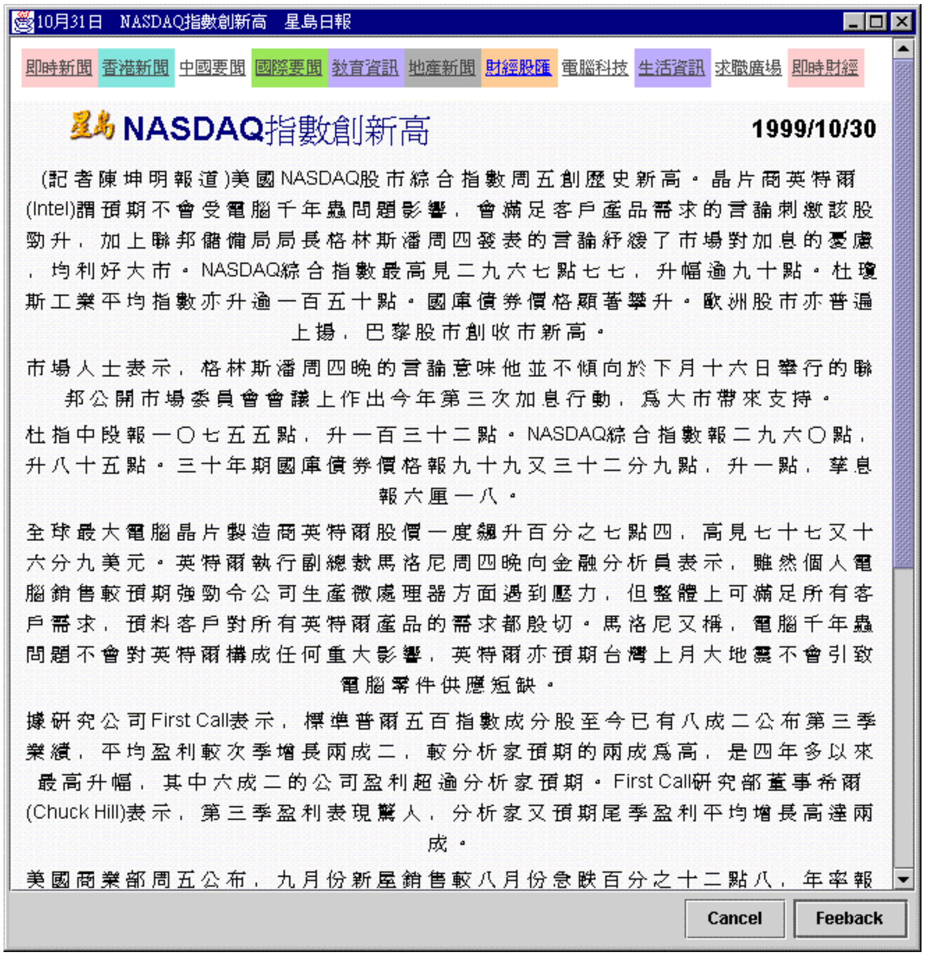
Rank	Score	Title	News	Date	Retrieved
1	100	黃金海岸遊艇會景詩情畫意	星島日報	1999年4月26日	□
2	100	黃星華：買樓好時機	東方日報	1999年4月26日	□
3	100	麗港城升價乏力	星島日報	1999年4月26日	□
4	100	魯賓辭職傳聞料將出現	明報	1999年4月26日	□
5	100	駿威街貨減股價挫	星島日報	1999年4月26日	□
6	100	駿威星港德祥落實配股	東方日報	1999年4月26日	□
7	100	香港經濟新發展機遇	香港商報	1999年4月26日	□
8	100	韻濤居公開發售138單位沽清	香港商報	1999年4月26日	□
9	100	銀行搶客勁減息	東方日報	1999年4月26日	□
10	100	銀禧賣地後僅錄一宗成交	明報	1999年4月26日	□
11	100	銀禧業主強硬市況呆滯	星島日報	1999年4月26日	□
12	100	鉅基閣呎價3440後日發售	明報	1999年4月26日	□
13	100	金盾純利增19% 符預期	明報	1999年4月26日	□
14	100	金朝陽勁蝕近五億	星島日報	1999年4月26日	□
15	100	金價應聲急挫三美元	東方日報	1999年4月26日	□
16	100	配股活動頻市受壓科技股當旺炒味濃	香港商報	1999年4月26日	□
17	100	過分超買遏升力	香港商報	1999年4月26日	□
18	100	逾萬法輪功信徒中南海示威	星島日報	1999年4月26日	□
19	100	近期住宅市場租售表現平均本月二手樓價上…	香港商報	1999年4月26日	□
20	100	趁外資流入紛紛減持	明報	1999年4月26日	□
21	100	買銅與夾板股	明報	1999年4月26日	□
22	100	財經快訊	香港商報	1999年4月26日	□

Figure 10. Result of Ranked Financial News on a Particular Day

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 five consecutive days. 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 every day. The performance is measured by precision of retrieval.

$$\begin{aligned}\text{Precision} &= \frac{\text{\# of relevant returned news articles}}{\text{\# of returned news articles}} \\ &= \frac{\text{\# of relevant returned news articles}}{20}\end{aligned} \tag{7}$$

Figure 12 shows the experimental results. For the first setup, in which only the 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 day, if there are not too many articles that are of interest to 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, in 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 helps 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 used, the performance is significantly better than using user profile or using feedback only on Day 1 and Day 2. The performance comes closer to that of using feedback only but is still significantly better than using profile only. The decrease in precision on Day 4 may only be due to the lower number of relevant documents on that day compared to other days as we observe the decrease of precision on Day 4 for the first setup.



NASDAQ指數創新高　1999/10/30

(記者陳坤明報道)美國NASDAQ股市綜合指數周五創歷史新高。晶片商英特爾(Intel)謂預期不會受電腦千年蟲問題影響，會滿足客戶產品需求的言論刺激該股勁升，加上聯邦儲備局局長格林斯潘周四發表的言論紓緩了市場對加息的憂慮，均利好大市。NASDAQ綜合指數最高見二九六七點七七，升幅逾九十點。杜瓊斯工業平均指數亦升逾一百五十點。國庫債券價格顯著攀升。歐洲股市亦普遍上揚，巴黎股市創收市新高。

市場人士表示，格林斯潘周四晚的言論意味他並不傾向於下月十六日舉行的聯邦公開市場委員會會議上作出今年第三次加息行動，為大市帶來支持。

杜指中段報一〇七五五點，升一百三十二點。NASDAQ綜合指數報二九六〇點，升八十五點。三十年期國庫債券價格報九十九又三十二分九點，升一點，孳息報六厘一八。

全球最大電腦晶片製造商英特爾股價一度飆升百分之七點四，高見七十七又十六分九美元。英特爾執行副總裁馬洛尼周四晚向金融分析員表示，雖然個人電腦銷售較預期強勁令公司生產微處理器方面遇到壓力，但整體上可滿足所有客戶需求，預料客戶對所有英特爾產品的需求都殷切。馬洛尼又稱，電腦千年蟲問題不會對英特爾構成任何重大影響，英特爾亦預期台灣上月大地震不會引致電腦零件供應短缺。

據研究公司First Call表示，標準普爾五百指數成分股至今已有八成二公布第三季業績，平均盈利較次季增長兩成二，較分析家預期的兩成為高，是四年多以來最高升幅，其中六成二的公司盈利超逾分析家預期。First Call研究部董事希爾(Chuck Hill)表示，第三季盈利表現驚人，分析家又預期尾季盈利平均增長高達兩成。

美國商業部周五公布，九月份新屋銷售較八月份急跌百分之十二點八，年率報

Figure 11. Agent Interface Displaying News Articles in News Browser

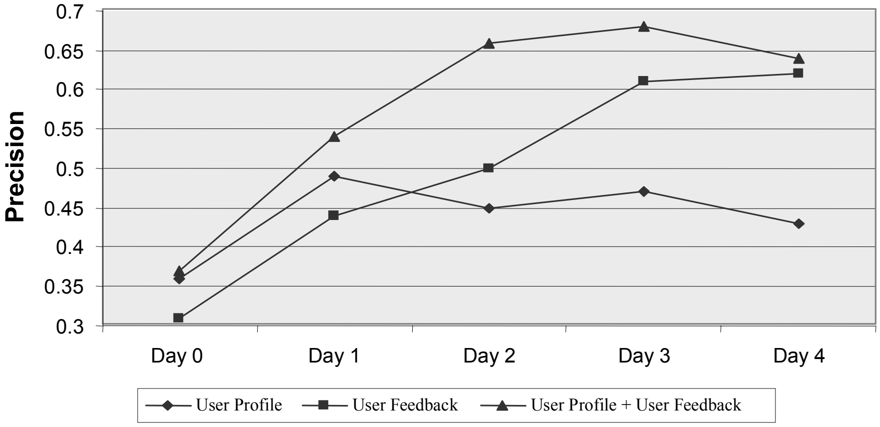


Figure 12. Experimental Results

4. CONCLUSION

We have presented an intelligent agent for retrieving Chinese financial news articles on the Web. User feedback and user profiles are utilized to learn 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 the Hong Kong 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 user preferences and indexing on behalf of users. We 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. User feedback helps in improving the retrieval performances continuously but the improvement stops after a period of time. Combining both user profiles and user feedback is significantly better than using either user profiles or user feedback only.

References

Armstrong, R., Freitage, D., Joachims, T., and Mitchell, T. "WebWatcher: A Learning Apprentice for the World Wide Web," *AAAI 1995 Spring Symposium Information Gathering from Heterogeneous, Distributed Environments*, Menlo Park, CA, 1995.

Chen, H., Chung, Y., Ramsey, M., and Yang, C. C. "A Smart Itsy Bitsy Spider for the Web," *Journal of the American Society for Information Science* (49:7), May 15, 1998, pp. 604-618.

DeBra, P., and Post, R. "Information Retrieval in the World Wide Web: Making Client-Based Searching Feasible," *Proceedings of the First International World Wide Web Conference*, Geneva, Switzerland, 1994.

Giles, C. L., Bollacker, K. D., and Lawrence, S. "CiteSeer: An Automatic Citation Indexing System," *Proceedings of the Third ACM Conference on Digital Libraries*, New York, 1998, pp. 89-98.

Kamba, T., Sakagami, H., and Koseki, Y. "Anatagonomy: A Personalized Newspaper on the World Wide Web," *International Journal of Human-Computer Studies* (46:6), June 1997, pp. 789-803.

Lieberman, H. "Autonomous Interface Agents," *Proceedings of the ACM Conference on Computers and Human Interface*, CHI-97, Atlanta, Georgia, March 1997.

Lieberman, H. "Letizia: An Agent That Assists Web Browsing," *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, Montreal, August 1995.

Pazzani, M., and Billsus, D. "Learning and Revising User Profiles: The Identification of Interesting Web Sites," *Machine Learning*, 27, Dordrecht, The Netherlands: Kluwer Academic Publishers, 1997, pp. 313-331.

Pazzani, M., Muramatsr, J., and Billsus, D. "Syskill & Webert: Identifying Interesting Web Sites," *Proceedings of the National Conference on Artificial Intelligence*, Portland, OR, 1997, pp. 54-61.

Pinkerton, B. "Finding What People Want: Experiences with the WebCrawler," *Proceedings of the Second International World Wide Web Conference*, Chicago, IL, October 17-20, 1994.

Salton, G., and Buckley, C. "Improving Retrieval Performance by Relevance Feedback," *Journal of the American Society for Information Science* (41:4), 1990, pp. 288-297.

Sakagami, H., and Kamba, T. "Learning Personal Preferences on Online Newspaper Articles from User Behaviors," *Sixth International on World Wide Web Conference*, Santa Clara, CA, April 7-11, 1997.

Yang, C. C., Luk, J. W. K., Yung, S. J., and Yen, J. "Combination and Boundary Detection Approaches on Chinese Indexing," *Journal of the American Society for Information Science* (51:4), 2000a, pp. 340-351.

Yang, C. C., Yen, J., and Chen, H. "Intelligent Internet Searching Agent Based on Hybrid Simulated Annealing," *Decision Support Systems*, 2000b.

Yang, C. C., Yen, J., Yung, S. K., and Chung, A. "Chinese Indexing with Mutual Information," *Proceedings of the First Asia Digital Library Workshop*, Hong Kong, August 6-7, 1998.