Appl Intell (2009) 30: 121–141
DOI 10.1007/s10489-007-0106-7

Personalized e-news monitoring agent system for tracking user-interested Chinese news events

Chih-Ming Chen · Chao-Yu Liu

Published online: 25 November 2007

Abstract Numerous paper-based newspapers have been transformed into a digital format and published on the Internet. Digital newspapers are gradually becoming a popular electronic media for conveying information immediately. Google developed a powerful news service, Google news alert, based on the Google news aggregator for tracking user-interested new events utilizing a keywords matching approach. However, this service only monitors and tracks news events using the keyword-matching scheme; consequently, the Google news alert retrieves many irrelevant news events and sends them to users. In other words, the current service cannot monitor news events via a specific news topic; although recall rate is high, the precision rate is low when tracking user-interested news events. Thus, this study presents a novel personalized e-news monitoring agent system that employs the topic-tracking-based approach, improving the flaw of the keyword-based approach, for tracking user-interested news events on Google News site. The proposed scheme simultaneously considers both similarities and the semantic relationships among news topics to track news events. Additionally, to further support the promotion of the accuracy rate in tracking user-interested Chinese news events, the Chinese word segmentation system ECScanner (An Extension Chinese Lexicon Scanner) with new word extension is proposed for the Chinese word segmentation process. Experimental results demonstrated that the proposed scheme, based on topic-based approach, is superior to the keyword-based approach used by Google news alert in terms of precision rate, and retains a high recall rate when tracking user-interested news events. Compared with the conventional Chinese word segmentation system CKIP (Chinese Knowledge Information Processing), experimental results also confirmed that using the proposed ECScanner with novel extension mechanism for new words improves the accuracy rate in tracking user-interested news events.

Keywords News events · News events monitoring agent system · Information retrieval · Intelligent agent

1 Introduction

The World Wide Web has become the predominant electronic media for information delivery. Traditional paper-based newspapers are expanding their services by providing on-line news on the World Wide Web to increase their competitiveness and profits. News documents frequently contain information that is useful and valuable when investigating social development trends, stock investments [1–7], and decision making analysis [8, 9]. Although many commercial portals provide immediate on-line news using a basic search mechanism and categorize news articles for users, few news web sites provide intelligent schemes for news services. Actually, reading or monitoring everyday news stories from the Internet is a difficult and time-consuming job for modern humans. Developing intelligent mechanisms for news services has practical needs.

To survey past studies, few studies have paid attention to exploring news articles for information retrieval

C.-M. Chen (✉)
Graduate Institute of Library, Information and Archival Studies,
National Chengchi University, No. 64, Sec. 2, ZhiNan Rd.,
Wenshan District, 116 Taipei, Taiwan
e-mail: chencm@nccu.edu.tw

C.-Y. Liu
Graduate Institute of Learning Technology, National Hualien
University of Education, No. 123, Hua-Hsi Rd., 970 Hualien,
Taiwan
e-mail: rank@enjust.com

or value-added information in recent years. Some related studies were briefly summarized and compared with this study herein. Shah and Elbahesh [10] proposed a clustering scheme for news topics based on similarity measures that provides categorized news events for user browsing. Maria and Silva [11] created a classification scheme for news topics to retrieve interested news themes using a support vector machine. Kurtz and Mostafa [12] proposed an approach for tracking user-interested news topics based on a keyword-matching technique and clustering scheme. Moreover, Lam and Cheung *et al.* [13] developed a matching algorithm that translates unknown English vocabularies into the corresponding Chinese characters. Allan and Papka *et al.* [14] proposed utilizing similarity measures to track the history of specific news events, and Lee *et al.* [15] proposed a novel scheme for constructing a fuzzy ontology based on a domain ontology and applied it to effectively summarize weather forecast news. These studies above-mentioned mainly focused on exploring the category of a news event by a classification or clustering scheme, but tracking the long-term developing tracks of a news story from the Internet has not been significantly concerned yet in these studies. The long-term developing tracks of a news story often contain richer information which is potentially valuable to be explored for value-added applications than a single news event.

Additionally, previous studies [1–9] have indicated that a relationship exists between news articles and stock prices. Predicting stock price movement based on news articles is an emerging topic in the data-mining field [16]. Wiithrich *et al.* [4, 6] proposed using information contained in articles published on the Internet to predict stock market prices. Peramunetilleke *et al.* [7] investigated how money market news headlines can be utilized to forecast intraday currency exchange rate movements. Furthermore, several studies utilized text documents for stock price predictions [3–6].

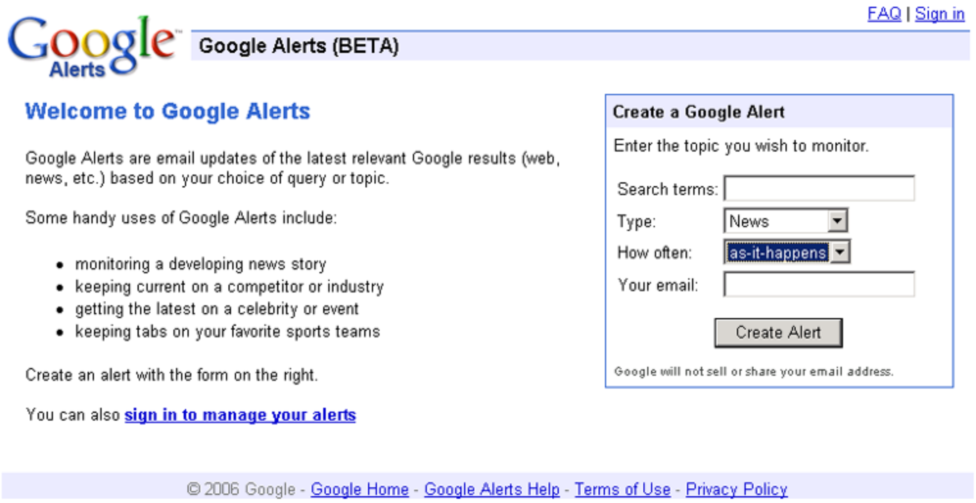
To monitor developing tracks of on-line e-news events effectively and efficiently, an aided e-news event monitoring agent system that assists individuals in obtaining user-interested news events immediately based on news topics is urgently needed. Google has developed a unique news service for monitoring the developing tracks of an online news story, Google news alert [17], based on the Google news aggregator to track user-interested news events using a keyword-matching scheme and send tracking results to users based on individual requirements via e-mail. However, this service only monitors and tracks news events using a keyword-matching scheme; consequently, many irrelevant news events are retrieved, resulting in a high recall rate and low precision rate. To improve the performance of Google news alert, this study primarily focuses on developing a personalized e-news monitoring agent system based on the proposed topic-tracking-based scheme for tracking user-interested Chinese news events.

Experimental results indicated that the proposed topic-tracking-based scheme has a better precision rate than the keyword-based approach used by Google news alert and retains a satisfactory recall rate while tracking user-interested news events. Moreover, Foo's study [18] indicated that an effective Chinese word segmentation process can increase the probability of a query-document match, thus enhancing the accuracy of information retrieval. Although the CKIP (Chinese Knowledge Information Processing) with new word guessing [19] is the most frequently used Chinese word segmentation system as well as won the first prize in the competition of the first international Chinese word segmentation bakeoff [20] among 6 participating systems for traditional Chinese texts in 2003, it is a general purpose Chinese word segmentation system developed by the top research institute Academia Sinica in Taiwan. Thus, it cannot segment domain-specific Chinese texts like financial news articles well. To promote further the accuracy rate of tracking user-interested news events, the Chinese word segmentation system CKIP is replaced by the proposed Chinese word segmentation system ECScanner that has a new word extension mechanism [21] to assist Chinese word segmentation when tracking user-interested news events. Experimental results show that ECScanner with new word extension mechanism is superior to CKIP in identifying meaningful Chinese words, clearly enhancing the accuracy of tracking user-interested news events.

The remainder of this paper is organized as follows. Problem description for tracking user-interested news events from Google news aggregator is first explained in Sect. 2. The system architecture of the proposed personalized e-news event monitoring agent system is detailed in Sect. 3. Section 4 proposes a novel Chinese word segmentation system ECScanner for supporting the accuracy promotion of tracking user-interested news events, and Sect. 5 explains the proposed two-phase scheme for tracking user-interested news events. Sections 6 and 7 present experimental results and discussion, respectively. Finally, the conclusion is stated in Sect. 8.

2 Problem description

Google news aggregator currently accesses nearly 10,000 news sources on the Internet via an automatic crawler program. The content from these sources is presented as news stories/categories in a searchable format on the web [22]. Without regard to political viewpoint or ideology, leading stories selected automatically by a computer algorithm are viewed as headlines on the Google news home page. Google news uses an automated process to group related headlines together into a news story/category, which, in some cases, enables people to access different viewpoints on the



Google Alerts (BETA)

Welcome to Google Alerts

Google Alerts are email updates of the latest relevant Google results (web, news, etc.) based on your choice of query or topic.

Some handy uses of Google Alerts include:

- monitoring a developing news story
- keeping current on a competitor or industry
- getting the latest on a celebrity or event
- keeping tabs on your favorite sports teams

Create an alert with the form on the right.

You can also sign in to manage your alerts

Create a Google Alert

Enter the topic you wish to monitor.

Search terms:
Type: News
How often: as-it-happens
Your email:

Google will not sell or share your email address.

© 2006 Google - Google Home - Google Alerts Help - Terms of Use - Privacy Policy

Fig. 1 The user interface of Google news alert for English language family [17]

same story/category [22]. News topics are updated continuously throughout the day and readers can view new stories/categories by checking the Google news website, subscribing to Google news alerts via email, or activating a Really Simple Syndication (RSS). Based on our evaluation, the automated process conducts a very high accuracy rate, even it does not lose in manual classification. The Google news service is currently tailored to 22 international audiences [22].

Google alerts currently offer three types of alerts: News, Web, and News and Web [17]. The Google news alert is an email message notifying a user of new articles appearing in the top 10 results of a Google news search [17]. When a user sets up a news alert, the frequency selected (daily, weekly, or as it happens) determines how frequently Google checks for news, not necessarily how often a user receives alerts. Google news alert provides different user interfaces to track news articles for various language families. Figure 1 displays the user interface for Google news alert for the English language family. To evaluate the drawbacks of Google news alert while tracking user-interested news events, a large amount of news titles randomly selected from Google news site was used as user queries to assess the accuracy rate of tracking user-interested news events. Based on our test results, this study determined that Google news alert can only set up short keywords when monitoring user-interested news events, and cannot set up an entire news event relating to the setting news event due to the keyword matching scheme for tracking news events. The keyword matching scheme easily gathers too many off-topic results. Google FAQ documents suggest that users can refine keywords to increase the accuracy of tracking user-interested news events according to advanced search results via heuristic tries [23]. The approach suggested by Google is clearly not a good solution.

Therefore, this study employed a topic-tracking-based scheme to improve the flaw of the keyword-based approach employed by Google news alert for tracking user-interested Chinese news events. This study mainly focused on a user must first specify his/her interested news topics by providing complete news titles and similarity thresholds through the user interface for determining user-interested news events, then the proposed news events monitoring agent assists the user to track the news events related to the user specified interested news event by the proposed two-phase scheme. In the study, the user-interested news events are defined as: a user feels valuable news events that he/she would pay attention to their long-term developing tracks in the future. In other words, the study aims at tracking the long-term developing tracks of a user specified interested news topic, not exploring user-interested news events based on users' news reading or browsing behavior. Therefore, providing complete title words of a news event is enough to track the news events relating to a user specified interested news event. Modeling users preferences for users' news reading or browsing behavior by certain quantitative user modeling technique is not the focused issue of this study. The following sections explain the proposed scheme in detail.

3 System architecture

This study presents a personalized news event monitoring agent system that contains both a news archive module and personalized news service module to track user-interested news events on the Google News site. Figure 2 presents the system architecture. The news archive module automatically archives news event metadata extracted using a metadata extraction process from the Google news aggregator via a crawler agent scheduled by a scheduling agent. These retrieved metadata, including the news story title, published medium, reporter, location, date, news body, and hyperlink to the original published medium, are stored in the news

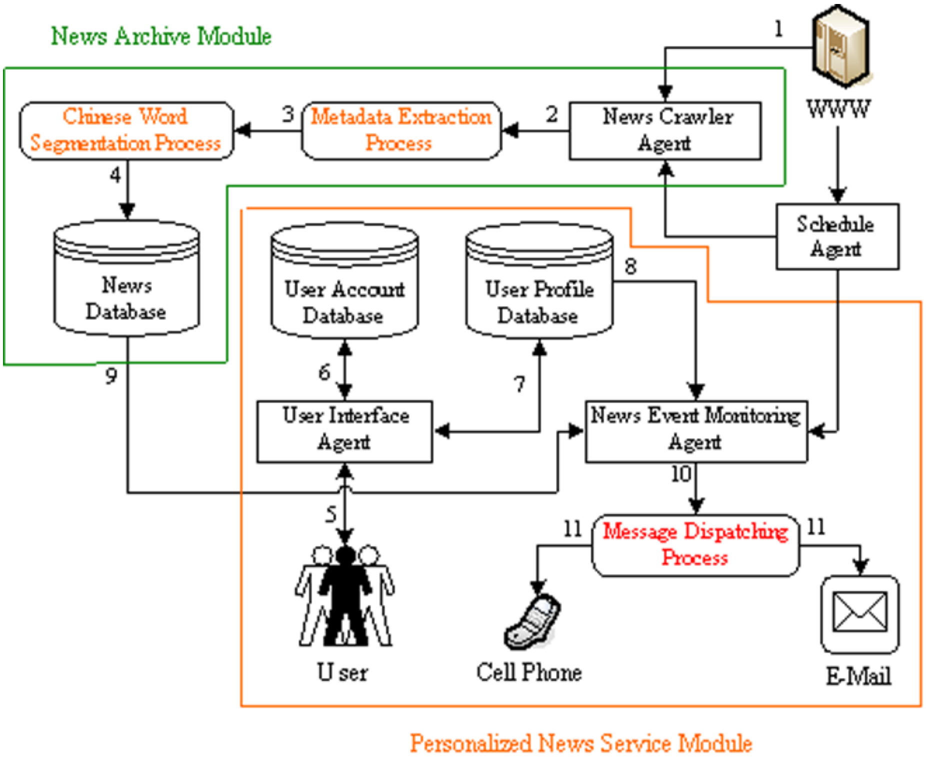


Fig. 2 System architecture of the personalized e-news event monitoring agent system

database. Among these seven retrieval new event metadata, news titles are further subjected to Chinese word segmentation process using CKIP, and the proposed ECScanner for utilizing to track user-interested Chinese news events. Furthermore, the personalized news service module primarily contains the news event monitoring agent that tracks and dispatches user-interested news events to individuals via a short message to cell phones or e-mail to mailboxes when user-interested news events appear on the Google News site. The user interface agent displays tracking results of news events based on individual user preferences in the user profile database. The user profile database records individual user preferences including the interested news titles, assigned threshold of similarity measure, dispatching frequency, and dispatching mode for tracking user-interested news events from Google News site. The user account database records registration information of individuals including accounts, passwords, e-mail boxes, and cell phone numbers.

3.1 System components

The system components of the proposed personalized news event monitoring agent system are introduced in detail in this section. A computing entity that performs user delegated tasks autonomously is defined as an intelligent agent herein [24].

3.1.1 News crawler agent

The news crawler agent automatically retrieves news events from the Google news aggregator according to running time established in the schedule agent. The extracted news events is then passed to the metadata extraction process for parsing important metadata, such as the news title, reporter, medium, date, location, body text, and hyperlink to the original published medium.

3.1.2 Schedule agent

The schedule agent is the control center and coordinates operations between the news crawler agent and the news event-monitoring agent in the proposed e-news event monitoring agent system. The schedule agent can fine tune news extraction strategies in the news crawler agent based on a detected Internet flow rate.

3.1.3 Metadata extraction process

The metadata extraction process extracts important news metadata based on news templates in Google news aggregator. As the news templates in the Google news aggregator are organized by HTML tags, the metadata extraction process identifies news metadata according to the characteristics of front and rear HTML tags with extracted metadata. However, observing the characteristics of HTML tags for extracting news metadata must be first performed by manpower, and then a software program is designed to successfully identify useful metadata from HTML documents based on regular expressions.

3.1.4 Chinese word segmentation process

The Chinese word segmentation process segments a Chinese news title into several separate Chinese words for evaluating the similarities between user-interested news events and news events in the news database gathered from Google News site. This process is very important since it affects the accuracy rate in tracking user-interested news events. Since news events frequently contain words that seldom appear in previous news events, a Chinese word segmentation system with a fixed word lexicon for the word segmentation process cannot satisfy the requirement of segmenting Chinese news titles. In this study, the Chinese word segmentation system with a new words' extension mechanism, ECScanner [21], is proposed to perform the Chinese word segmentation process. This system is discusses in detail in Sect. 4.2.

3.1.5 User interface agent

The user interface agent is a user-friendly interface for setting detailed conditions for tracking user-interested news events and displays tracking results. Moreover, the user interface agent also provides an interface where a user can directly observe his/her tracking news when he/she has logged in the system. This agent can autonomously coordinate with the news event monitoring agent based on the user profile database for tracking user-interested news events.

3.1.6 News event monitoring agent

The news event-monitoring agent tracks whether user-interested news events are published on the Google new site. When user-interested news events are identified, the news events agent sends a message to the message dispatching process for dispatching the news events to individual users. The detailed scheme for implementing the news events monitoring agent for tracking user-interested news events is introduced in Sect. 5.

3.1.7 Message dispatching process

The message dispatching process sends the identified news events to individual users via a short message to cell phones or e-mail to mailboxes. If a user-interested news event has already been dispatched to the user, the proposed system will disable a dispatching flag to avoid repeatedly dispatching the news event. Therefore, only the new news is dispatched to a user.

3.2 System operation procedure

The following is the system operation procedure for the proposed news event monitoring agent (Fig. 2).

Step 1. The schedule agent senses the Internet environment to coordinate the news crawler agent for extracting required news events from the assigned news sites that archive news events.

Step 2. News events with an HTML format retrieved from the news events crawler agent are then subjected to the metadata extraction process. In this process, all redundant tags in retrieved news events are filtered out to preserve useful tags for extracting significant metadata, such as news story title, reporter, media, location, date, news body, and hyperlink to the original published medium.

Step 3. News story titles are then subjected to the Chinese word segmentation process by CKIP or the proposed ECScanner, thereby obtaining separate Chinese phrases for tracking user-interested news events using the proposed two-phase scheme based on a vector space model.

Step 4. These separated Chinese phrases obtained from news titles are stored in the news database.

Step 5. Users log onto the personalized e-news event monitoring agent system using a legal account for tracking interested news events.

Step 6. The system identifies a user account to determine whether a user can log onto this system.

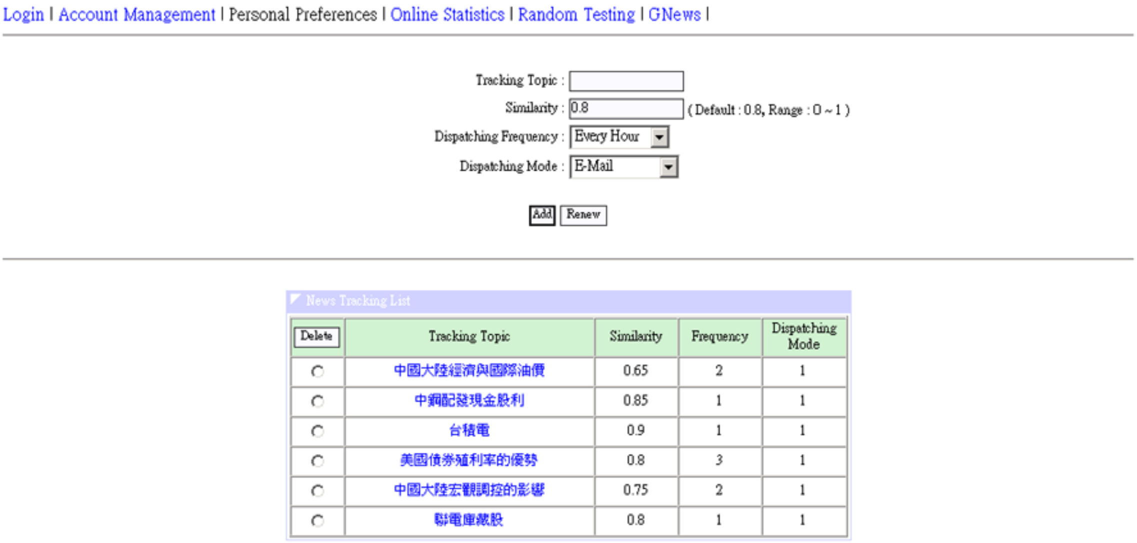
Step 7. If a user has a legal account, the system then retrieves the user's preferences to the user interface agent for modification or adding new event topics to be tracked. Figure 3 presents the user interface utilized when setting user preferences. When a user finishes modifying or adding news topics, the system uploads the new user preferences to the user profile database. In this study, a user can specify his/her preferences including interested news topics, similarity threshold for determining user-interested news events, news dispatching frequency, and news dispatching mode through the user interface. Currently, the importance of each user-interested news event is served as equivalent.

Step 8. The news event monitoring agent tracks news events according to user preferences obtained from the user profile database.

Step 9. User-interested news events are retrieved from the news database based on user preferences in the user profile database.

Step 10. Retrieved user-interested news events are sent to the message dispatching process.

Step 11. The system dispatches the retrieved user-interested news events via a short message to individual cell phones or e-mail to individual mailboxes. Figure 4 presents results of the tracking news events dispatched by e-mail. The dispatching messages through cell phone or e-mail include news title, hyperlink to the original news medium, news


Login | Account Management | Personal Preferences | Online Statistics | Random Testing | GNews |
Add
Renew
News Tracking List
Delete

Tracking Topic :

Similarity : 0.8 (Default : 0.8, Range : 0 ~ 1)

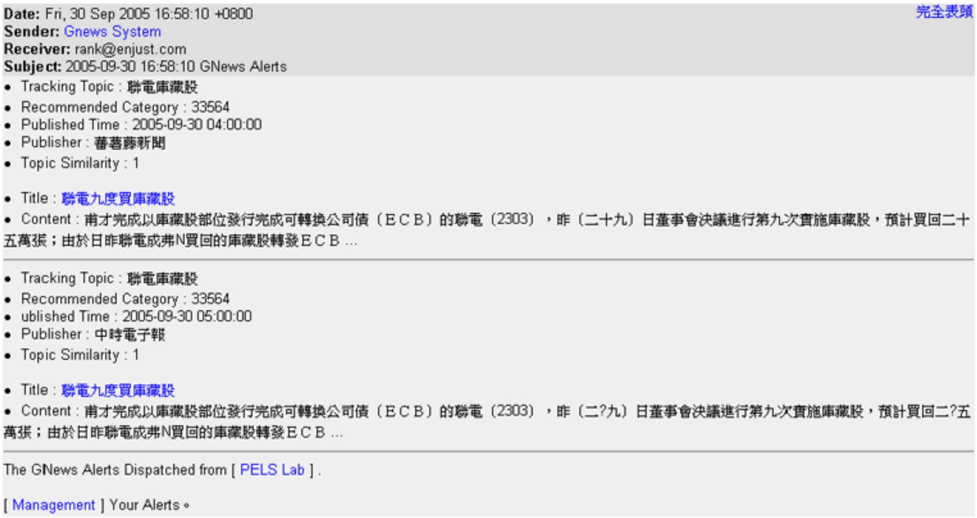
Dispatching Frequency : Every Hour

Dispatching Mode : E-Mail

Delete	Tracking Topic	Similarity	Frequency	Dispatching Mode
○	中國大陸經濟與國際油價	0.65	2	1
○	中鋼配發現金股利	0.85	1	1
○	台積電	0.9	1	1
○	美國債券殖利率的優勢	0.8	3	1
○	中國大陸宏觀調控的影響	0.75	2	1
○	聯電庫藏股	0.8	1	1

Fig. 3 The user interface for setting user's preference

Fig. 4 The dispatching results of the tracking user-interested news events by e-mail



Date: Fri, 30 Sep 2005 16:58:10 +0800
Sender: Gnews System
Receiver: rank@enjust.com
Subject: 2005-09-30 16:58:10 GNews Alerts

- Tracking Topic : 聯電庫藏股
- Recommended Category : 33564
- Published Time : 2005-09-30 04:00:00
- Publisher : 蕃薯藤新聞
- Topic Similarity : 1

- Title : 聯電九度買庫藏股
- Content : 甫才完成以庫藏股部位發行完成可轉換公司債（ＥＣＢ）的聯電（2303），昨（二十九）日董事會決議進行第九次實施庫藏股，預計買回二十五萬張；由於日昨聯電成弗N買回的庫藏股轉發ＥＣＢ ...

- Tracking Topic : 聯電庫藏股
- Recommended Category : 33564
- ublished Time : 2005-09-30 05:00:00
- Publisher : 中時電子報
- Topic Similarity : 1

- Title : 聯電九度買庫藏股
- Content : 甫才完成以庫藏股部位發行完成可轉換公司債（ＥＣＢ）的聯電（2303），昨（二?九）日董事會決議進行第九次實施庫藏股，預計買回二?五萬張；由於日昨聯電成弗N買回的庫藏股轉發ＥＣＢ ...

The GNews Alerts Dispatched from [PELS Lab] .

[Management] Your Alerts。

publisher, published date of news, and news body of the first paragraph.

4 The proposed Chinese word segmentation system for supporting the accuracy promotion of tracking user-interested news events

This section presents a word segmentation system with extension of news words for the accuracy promotion of tracking user-interested news events. First of all, the importance of word segmentation process for Chinese texts is addressed in Sect. 4.1. Section 4.2 details the proposed Chinese word segmentation system ECScanner with extension of new words.

4.1 Chinese word segmentation

Identifying English or the other western languages texts into distinct words is natural and trivial task. By contrast, it is a very challenge and difficult task for Chinese texts, since Chinese texts consist of a string of ideographic characters without any blanks to mark word boundaries between words except for punctuation signs at the end of each sentence, and occasional commas within sentences [18]. Therefore, it is well known that there are two major difficulties in Chinese word segmentation. One is resolving the ambiguous segmentation, and another is identifying unknown words [20]. However, the word segmentation is a necessary step in processing Chinese texts, such as machine translation, Chinese text mining and information retrieval. To survey the past studies [25, 26], developing Chinese word segmentation techniques mainly focused on three approaches including the lexicon-based identification scheme, statistics-based word identification scheme, and hybrid word identification scheme. The basic technique for identifying distinct words from Chinese texts is based on the lexicon-based identification scheme [27], which performs word segmentation process using string matching algorithms supported by a well prepared lexicon with sufficient amount of lexical entries which covers all of the Chinese words as possible. However, such a large lexicon is difficult to be constructed or maintained by manpower since the set of words is open-ended. Therefore, many used words in Chinese texts for word segmentation are often out-of-lexicon words due to insufficient amount of lexical entries so that the accuracy of Chinese word segmentation is degraded. The extraction of new words becomes a key technology for the lexicon-based Chinese word segmentation systems [20].

The CKIP [19] was designed to perform word segmentation for Chinese documents in general domains. However, CKIP cannot handle Chinese news titles or articles well in some particular domains such as the financial news domain. For example, the CKIP conducted Chinese word segmentation process for the financial news title "聯電股價走勢強勁盤中完成填權(UMC stock price is rising and price recovery finished during trading)" as "聯電(UMC)/股價(Stock Price)/走勢(Trend)/強勁(Rising)/盤中(Trading)/完成(Finish)." The segmentation result suggests that the word "填權(Price Recovery)" is not viewed as a meaningful word because the current word lexicon in CKIP did not contain the word "填權(Price Recovery)." Except adopting new word guessing scheme, maintaining the word lexicon manually is a major routine task for the promotion of Chinese word segmentation capability in CKIP [19]. However, maintaining a Chinese word lexicon by human power is a time-consuming and inefficient job. In particular, Chinese word segmentation systems with fixed lexicon are very difficult to handle the contents that vary highly with time well, such as Web or news texts, because they generate new words frequently. A Chinese word segmentation system with new word extension can overcome the above-mentioned problem as well as provide benefit in terms of enhancing the accuracy rate when tracking user-interested news events.

4.2 The proposed Chinese word segmentation system with extension of new words

To solve the problem of new words occurring frequently in news articles, this study presents a uniformity-based new word discovery scheme for Chinese news titles to enhance the Chinese word segmentation system and accuracy rate in tracking user-interested news events. Although inducing the appearance opportunity and position of new words from the combination of Chinese linguistic words is difficult, meaningful new words in news articles always have a high frequency of appearing in most news articles [28–33]. If a word is meaningful and can be considered a new word in Google news titles, then it should appear in numerous news events in the same news story/category, rather than being concentrated in a few news events in a news story/category [34]. Based on the property, this study proposes a uniformity-based new word discovery approach that can assist the Chinese word segmentation system CScanner [35] in extending its Chinese word lexicon automatically, thus speeding up its Chinese word segmentation capability. CScanner, an open-source Chinese lexical scanner based on two variants of the maximum matching heuristic of word identification, was developed for the Chinese word segmentation process. In this study, the Chinese word lexicon in CScanner was improved using the proposed new word extension mechanism. The CScanner combined with the new word extension mechanism is called ECScanner [21] herein. The ECScanner is currently published at http://dlll.nccu.edu.tw/~rank/ecscanner/ and opens to provide Chinese word segmentation serviced by the Simple Object Access Protocol (SOAP) web service mechanism.

Suppose a Chinese news title is composed of n separated words, the number of combinations of word segmentation from 1-gram to n-grams can be computed as

$$CN_i = \sum_{i=1}^{n} (n - i + 1) \tag{1}$$

where CN_i represents the number of combinations of i-gram in a Chinese news title, n is the total number of separated Chinese words in a Chinese news title.

All combinations of word segmentation from 1-gram to n-grams for a Chinese news title composed of n separated words are considered as possible new words, the employed uniformity measure is then utilized to evaluate their qualities for new word discovery [34]. The uniformity of possible new word t contained in news category c can be measured as

$$U_{ct} = -\sum_{d=1}^{n} p_{td} \log p_{td}; \qquad p_{td} = \frac{tf_{td}}{\sum_{d=1}^{n} tf_{td}} \tag{2}$$

where p_{td} is the probability that the possible new word t occurs in the news event d of news category c, tf_{td} is the word frequency of the possible new word t in the news event d, and n is the total number of news events categorized in the news category c.

In (2), when computing uniformity of a possible new word extracted from a Google news title exceeds an assigned uniformity threshold, this Chinese word can then be viewed as a candidate new word conveying meaningful information related to the news category. Furthermore, the candidate new word must be compared with the original Chinese word lexicon to determine whether the candidate word is a new word. Candidate new words not contained in the original word lexicon are considered new words in this study. Since the original Chinese lexicon has been stored in a relational database in the employed Chinese word segmentation system, adopting SQL query for comparing candidate new words with original Chinese lexicon is the simplest method. However, the executing performance will gradually decline with fast growth of archiving news events if this method was employed. To speed up the executing performance of comparing candidate new words with original Chinese lexicon, uploading the original Chinese lexicon into memory based on various gram words was adopted to effectively promoting executing performance in the study. That is, the strategy of searching relational database is replaced by the strategy of searching memory.

Moreover, subsets of discovered new words are frequently in the original word lexicon. For example, the discovered quadgram of "晶圓代工(Semiconductor Foundry)" can been segmented into the two bigrams of "晶圓(Semiconductor)" and "代工(Foundry)". Obviously, these two bigrams are subsets of the discovered quadgram of "晶圓代工(Semiconductor Foundry)". If the two bigrams of "晶圓(Semiconductor)" and "代工(Foundry)" are in the original word lexicon, how to handle this problem is an issue for the proposed new word discovery approach. In the proposed scheme, the two bigrams of "晶圓(Semiconductor)" and "代工(Foundry)" are replaced by the quadgram of "晶圓代工 (Semiconductor Foundry)" because the quadgram usually generates more abundant semantics than bigrams. However, when the number of news titles in the same news category contain the two bigrams "晶圓(Semiconductor)" and "代工(Foundry)" exceeds an assigned percentage threshold, the proposed scheme simultaneously preserves the discovered quadgram of "晶圓代工(Semiconductor Foundry)" and these two bigrams "晶圓(Semiconductor)" and "代工(Foundry)" in the word lexicon. The proposed new word discovery approach is summarized as follows.

Step 1. Generate all combinations of segmented words from 1-gram to n-grams for a Chinese news title with n separated Chinese words as possible new words, and compute their uniformity values using (2).

Step 2. Preserve possible new words with corresponding uniformity values exceeding an assigned threshold as candidate new words.

Step 3. Generate new words by eliminating candidate new words in the original word lexicon.

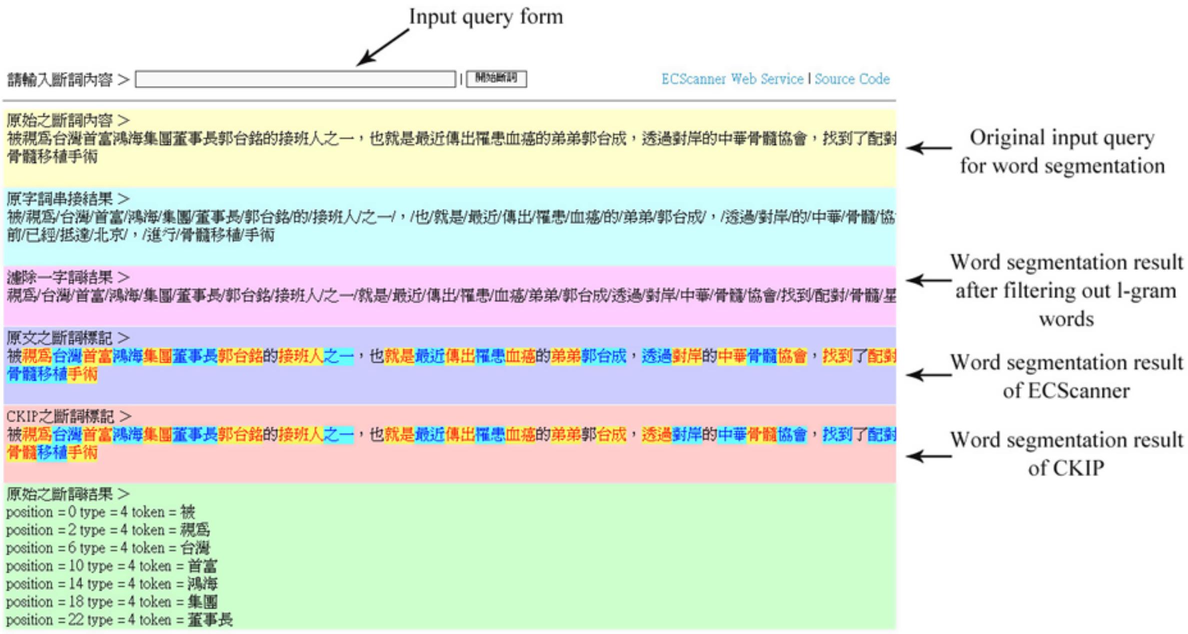
Step 4. Determine whether subsets of discovered new words are contained in the original word lexicon to determine which strategy should be used for new word generation.

Case 1. When subsets of the discovered new words are not in the original word lexicon, then all word segmentation outcomes are considered new words and recommended to linguistic experts for final confirmation.

Case 2. When subsets of discovered new words are in the original word lexicon, then few-grams are replaced by multi-grams based on semantic consideration except the number of news titles categorized in the same news category containing the few-grams is over an assigned percentage threshold.

Figure 5 shows the implemented ECScanner with new word discovery mechanism, which provides the Chinese word segmentation service by the SOAP web service. Additionally, the system provides simultaneously Chinese word segmentation results from CKIP with an unknown word guessing scheme and the proposed ECScanner with a uniformity-based new word discovery scheme. Experimental results discussed later prove that the performance of Chinese word segmentation of the proposed ECScanner with new word discovery mechanism is superior to CKIP with new word guessing mechanism.

In addition, new word discovery schemes have difficulty producing completely correct new words. In other words, when the lexicon in a Chinese word segmentation system completely accepts all new words discovered from a new word discovery scheme, the precision rate of the Chinese word segmentation scheme will gradually decline with time. A relatively better strategy is to develop a new word management system that provides a friendly interface for linguistic experts to assess whether the discovered new words should be formally accepted as new words or rejected as un-meaningful words through an administrator interface. In our experiments, two linguistic experts from Department of Chinese Literature of National Chengchi University were invited to assess the discovered new words [36]. Figure 6 presents the user interface of the implemented new word management system, which has three operational modes—accept, reject or recommend—to assist linguistic experts in determining whether candidate words discovered by the proposed new word discovery scheme should be accepted as formal new words or rejected as un-meaningful words. After the new words are confirmed by linguistic experts, the Chinese word lexicon in ECScanner is immediately updated based on the planning execution time managed by the schedule agent. To promote the efficiency of maintaining new words, all reject new words confirmed by linguistic experts will not be recommended as likely new words again in the proposed system. This mechanism will gradually reduce the working load of linguistic experts with time.


Input query form
Original input query for word segmentation
Word segmentation result after filtering out l-gram words
Word segmentation result of ECScanner
Word segmentation result of CKIP

請輸入斷詞內容 > | 開始斷詞 ECScanner Web Service | Source Code

原始之斷詞內容 >
被視爲台灣首富鴻海集團董事長郭台銘的接班人之一，也就是最近傳出罹患血癌的弟弟郭台成，透過對岸的中華骨髓協會，找到了配對
骨髓移植手術

原字詞串接結果 >
被/視爲/台灣/首富/鴻海/集團/董事長/郭台銘/的/接班人/之一/，/也/就是/最近/傳出/罹患/血癌/的/弟弟/郭台成/，/透過/對岸/的/中華/骨髓/協
前/已經/抵達/北京/，/進亍/骨髓移植/手術

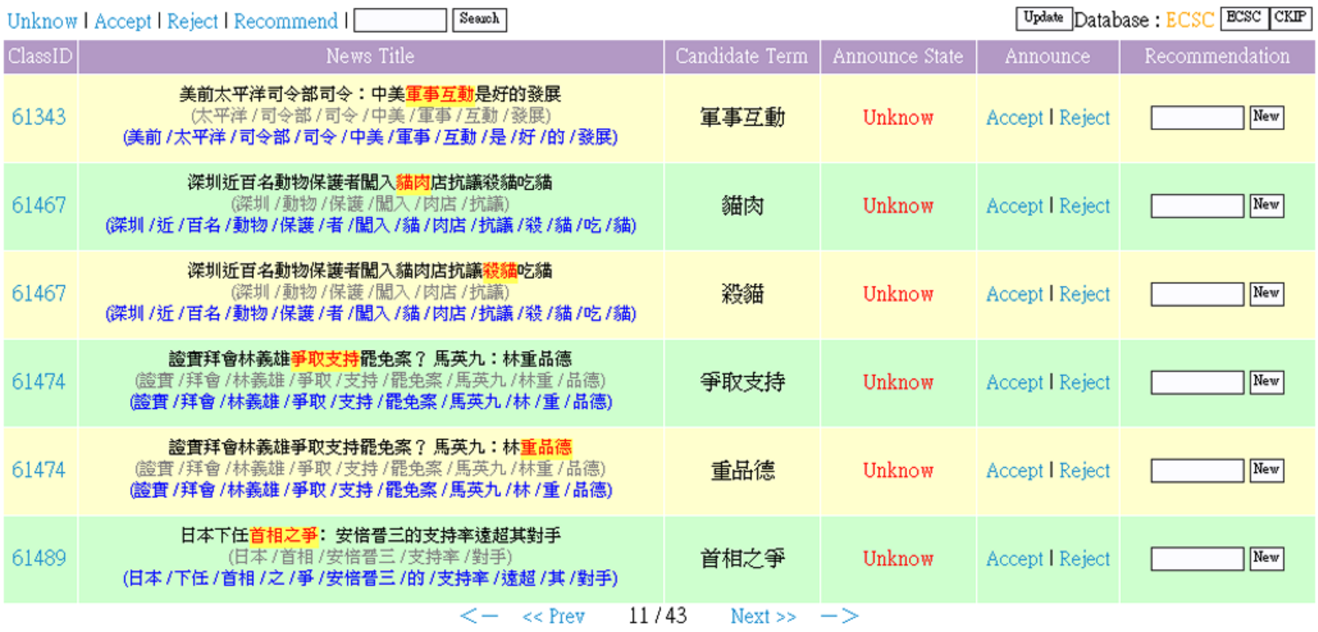
濾除一字詞結果 >
視爲/台灣/首富/鴻海/集團/董事長/郭台銘/接班人/之一/就是/最近/傳出/罹患/血癌/弟弟/郭台成/透過/對岸/中華/骨髓/協會/找到/配對/骨髓/星

原文之斷詞標記 >
被視爲台灣首富鴻海集團董事長郭台銘的接班人之一，也就是最近傳出罹患血癌的弟弟郭台成，透過對岸的中華骨髓協會，找到了配對
骨髓移植手術

CKIP之斷詞標記 >
被視爲台灣首富鴻海集團董事長郭台銘的接班人之一，也就是最近傳出罹患血癌的弟弟郭台成，透過對岸的中華骨髓協會，找到了配對
骨髓移植手術

原始之斷詞結果 >
position = 0 type = 4 token = 被
position = 2 type = 4 token = 視爲
position = 6 type = 4 token = 台灣
position = 10 type = 4 token = 首富
position = 14 type = 4 token = 鴻海
position = 18 type = 4 token = 集團
position = 22 type = 4 token = 董事長

Fig. 5 The Chinese word segmentation system ECScanner



Unknow | Accept | Reject | Recommend | Search Update Database : ECSC ECSC CKIP

ClassID	News Title	Candidate Term	Announce State	Announce	Recommendation
61343	美前太平洋司令部司令：中美軍事互動是好的發展 (太平洋 / 司令部 / 司令 / 中美 / 軍事 / 互動 / 發展) (美前 / 太平洋 / 司令部 / 司令 / 中美 / 軍事 / 互動 / 是 / 好 / 的 / 發展)	軍事互動	Unknow	Accept | Reject	New
61467	深圳近百名動物保護者闖入貓肉店抗議殺貓吃貓 (深圳 / 動物 / 保護 / 闖入 / 肉店 / 抗議) (深圳 / 近 / 百名 / 動物 / 保護 / 者 / 闖入 / 貓 / 肉店 / 抗議 / 殺 / 貓 / 吃 / 貓)	貓肉	Unknow	Accept | Reject	New
61467	深圳近百名動物保護者闖入貓肉店抗議殺貓吃貓 (深圳 / 動物 / 保護 / 闖入 / 肉店 / 抗議) (深圳 / 近 / 百名 / 動物 / 保護 / 者 / 闖入 / 貓 / 肉店 / 抗議 / 殺 / 貓 / 吃 / 貓)	殺貓	Unknow	Accept | Reject	New
61474	諮實拜會林義雄爭取支持罷免案？ 馬英九：林重品德 (諮實 / 拜會 / 林義雄 / 爭取 / 支持 / 罷免案 / 馬英九 / 林重 / 品德) (諮實 / 拜會 / 林義雄 / 爭取 / 支持 / 罷免案 / 馬英九 / 林 / 重 / 品德)	爭取支持	Unknow	Accept | Reject	New
61474	諮實拜會林義雄爭取支持罷免案？ 馬英九：林重品德 (諮實 / 拜會 / 林義雄 / 爭取 / 支持 / 罷免案 / 馬英九 / 林重 / 品德) (諮實 / 拜會 / 林義雄 / 爭取 / 支持 / 罷免案 / 馬英九 / 林 / 重 / 品德)	重品德	Unknow	Accept | Reject	New
61489	日本下任首相之爭：安倍晉三的支持率遠超其對手 (日本 / 首相 / 安倍晉三 / 支持率 / 對手) (日本 / 下任 / 首相 / 之 / 爭 / 安倍晉三 / 的 / 支持率 / 遠超 / 其 / 對手)	首相之爭	Unknow	Accept | Reject	New

<— << Prev 11 / 43 Next >> —>

Fig. 6 The news words management interface for determining whether the candidate words should be accepted as formal new words or rejected as un-meaningful words by linguistic experts

5 The proposed two-phase scheme for tracking user-interested news events

This section details how to modify the cosine measure to propose the two-phase scheme for tracking user-interested news events.

5.1 The cosine measure for evaluating the similarity between two news events

The news database retrieved from the Internet by the news crawler agent contains a large number of news events and grows very rapidly. The news crawler agent actively extracts everyday news events on the Google news aggregator according to running time scheduled by the schedule agent. To improve the computational efficiency and accuracy in tracking user-interested news events, the news event monitoring agent system employed the proposed two-phase scheme based on the cosine measure [34] for determining whether the user-interested news events are in the news database, then the user-interested news events are dispatched to individual users. In the first phase, the cosine measure was first applied to identify the similarity degrees between user-interested news events and news stories/categories classified together by the Google news aggregator, which contains news events reported by various news media. Consequently, the proposed scheme avoids comparing user-interested news events with each news event stored in the news database, thus promoting the computational efficiency while calculating similarity degrees for tracking user-interested news events in the first phase. A stronger property than the other similarity measures is that the cosine similarity does not depend on the length [34]. This allows documents with the same composition, but different totals to be treated identically which makes this the most popular measure for text documents. Also, due to this property, samples can be normalized to the unit sphere for more efficient processing [37]. Next, the first phase for tracking user-interested news events is presented as follows.

Suppose that the union term set obtained from the news event A and news events categorized in the qth news category contains a total of k linguistic terms extracted from news titles using a Chinese word segmentation system. The similarity degree between the news event A with the qth news story/category measured by the cosine measure can be formulated as

$$Sim(A, C_q) = \frac{\sum_{i=1}^{k} t_{Ai} t_{C_q i}}{\sqrt{\sum_{i=1}^{k} t_{Ai}^2 \sum_{i=1}^{k} t_{C_q i}^2}} \tag{3}$$

where $A = (t_{A1}, t_{A2}, \ldots, t_{Ai}, \ldots, t_{Ak})$ and $C_q = (t_{C_q 1}, t_{C_q 2}, \ldots, t_{C_q i}, \ldots, t_{C_q k})$ represent respectively the term vectors of the user-interested news event A and the qth news story/category, t_{Ai} and $t_{C_q i}$ stand for the binary term weights of the ith term in the user-interested news event A and the qth news story/category, respectively.

In (3), the corresponding binary term weight in the term vector is assigned as 1 when the term appears in the title of a news event; otherwise, the term weight is assigned as 0. This is because all term weights measured by term frequency have the same weight when two term vectors computing similarity by cosine measure contains no duplicate terms to each other. In this case, frequency information will degenerate into binary information. In other words, this situation is equal to computing similarity between two term vectors by binary term weights. In this study, news title was only considered information to track user-interested news events. Generally, news titles appear infrequently duplicate terms, this is the main reason that binary term weights were adopted to represent the term vectors of user-interested news event and news story/category while computing similarities between a user-interested news event with the news events stored in the news database by cosine measure. In the first phase, if similarities between a user-interested news event and the news stories/categories stored in the news database exceed an assigned threshold of the cosine measure, then these news stories/categories are preserved as candidate news stories/categories, and the second phase begins for further processing; otherwise, the news stories/categories are filtered out from the news stories/categories. In the second phase, the proposed scheme utilized the average modified cosine measure for each news event in the candidate news stories/categories to measure further the similarities between the user-interested news event and the candidate news stories/categories. In this work, the cosine measure is also employed to measure the similarity between two news events.

Suppose that the union terms' set obtained from the user-interested new event A and news event B contains a total of k linguistic terms extracted from news titles by a Chinese word segmentation system. The similarity degree between two news events measured by the cosine measure can be formulated as

$$Sim(A, B) = \frac{\sum_{i=1}^{k} t_{Ai} t_{Bi}}{\sqrt{\sum_{i=1}^{k} t_{Ai}^2 \sum_{i=1}^{k} t_{Bi}^2}} \tag{4}$$

where $A = (t_{A1}, t_{A2}, \ldots, t_{Ai}, \ldots, t_{Ak})$ and $B = (t_{B1}, t_{B2}, \ldots, t_{Bi}, \ldots, t_{Bk})$ represent respectively the term vectors of the user-interested news event A and the news event B categorized in a news category which passes an assigned threshold of similarity degree in the first phase, t_{Ai} and t_{Bi} stand for the binary term weights of the ith term in the user-interested news event A and the news event B, respectively.

In (4), the corresponding binary term weight in the term vector is assigned 1 when the term appears in the title of news event; otherwise, the term weight is 0. Next, this study gives two examples to illustrate why and how the cosine measure is employed to compute the similarity between two news events with binary term weights, not the distance measure employed. Suppose that the news events A and B are identified two and four separated Chinese terms from the news titles after performing Chinese word segmentation process, respectively. Further suppose that the terms t_2 and t_3 appear in the news title A, and the terms t_1, t_2, t_4, and t_5 appear in the news title B. Based on the proposed computing formula of similarity degree between two news events measured by the cosine measure, the union terms' set obtained from the new event A and news event B contains a total of 5 linguistic terms. That is, the term set $\{t_1, t_2, t_3, t_4, t_5\}$ is employed to represent each news event according to the corresponding term that appears in each news event. Thus, the news events A and B are represented as term vectors $(0, 1, 1, 0, 0)$ and $(1, 1, 0, 1, 1)$ based on the union terms' set obtained from the new event A and news event B. The similarity between the news events A and B is equal to $\frac{1}{\sqrt{8}}$ based on the proposed cosine measure, but the similarity between the news events A and B is equal to $\frac{1}{5}$ based on the distance measure.

Taking another example, suppose that the term t_1 appears in the news title C, and the terms t_1, t_2, t_3, t_4, and t_5 appear in the news title D. In this case, the term set $\{t_1, t_2, t_3, t_4, t_5\}$ is employed to represent each news event according to the corresponding term that appears in each news event. Thus, the news events C and D are represented as term vectors $(1, 0, 0, 0, 0)$ and $(1, 1, 1, 1, 1)$ based on the union terms' set obtained from the new event C and news event D. The similarity between the news events C and D is equal to $\frac{1}{\sqrt{5}}$ based on the proposed cosine measure, but the similarity between the news events C and D is also equal to $\frac{1}{5}$ based on the distance measure. In the two cases mentioned-above, it is very obvious that the distance measure cannot identify the difference of both the cases. By contrast, the proposed cosine measure can identify that the second case is more similar than the first case. The reason is that the distance measure only considers the occupied proportion of the same terms in the union terms' set while measuring similarity between two news events, but it ignores the occupied proportion of differentiated terms to each other. In other words, the distance measure is inappropriate to measure two term vectors, which have different term dimensions like the solving problem in the study.

5.2 The modified cosine measure with term semantics for evaluating the similarity between two news events

Moreover, considering only the similarity degree determined by the cosine measure cannot completely reveal a term's semantics. For example, the similarity degree of news titles "散戶買超台股(Individual investors overbuy Taiwan stocks)外資賣超(Foreign invertors oversell Taiwan stocks)", and "散戶賣超(Individual investors oversell Taiwan stocks)外資買超台股(Foreign invertors overbuy Taiwan stocks)" measured by cosine measure is equal to 1 since both the news titles include the same set of Chinese linguistic terms, i.e., 散戶(Individual investors)/買超(Overbuy)/台股(Taiwan stocks)/外資(Foreign investors)/賣超(Oversell), after the Chinese word segmentation process. This similarity degree evaluated by cosine measure implies that both news events are completely identical in terms of semantics, but have inverse meanings from the perspective of natural language. To consider a term's semantics for promoting the accuracy when tracking user-interested news events, the proposed hamming distance was applied to modify the cosine measure for measuring the similarity degree between two news events with increased precision. The proposed hamming distance for both the news events can be computed using the following formula:

$$H_{AB} = \sum_{i=1}^{m} |L(T_{Ai}) - L(T_{Bi})| \tag{5}$$

where $L(T_{Ai})$ and $L(T_{Bi})$ represents respectively the transformed location index values of the ith term that appears simultaneously in the titles of both the user-interested news event A and news event B, m is the total number of terms that appear simultaneously in the titles of both the user-interested news event A and news event B.

Next, this study gives an example explaining how to compute the hamming distance between two news events. Suppose that the news event A contains four terms after performing the Chinese word segmentation process, and the term set of news event A is represented as $T(A) = \{t_1, t_2, t_3, t_4\}$. To compute hamming distance, term set $T(A)$ must be transformed into a location vector based on the order in which terms appear. For example, the transformed location vector of term set $T(A)$ is represented as $L(A) = (1, 2, 3, 4)$ for the corresponding term set $T(A) = \{t_1, t_2, t_3, t_4\}$. Moreover, suppose that the term sets of another two news events B and C are $T(B) = \{t_1, t_3, t_5\}$ and $T(C) = \{t_1, t_5, t_3\}$, respectively. Similarly, the transformed location vectors for both B and C are $L(B) = (1, 2, 3)$ for term sets $T(B) = \{t_1, t_3, t_5\}$, as well as $L(C) = (1, 2, 3)$ for term set $T(C) = \{t_1, t_5, t_3\}$. Therefore, the hamming distances between news event A with the news events B and C can be computed as

$$H_{AB} = |1 - 1| + |3 - 2| = 1,$$

$$H_{AC} = |1 - 1| + |3 - 3| = 0.$$

Based on computed results for hamming distances, news events A and C have more similar semantics than news events A and B as hamming distance H_{AB} is larger than hamming distance H_{AC}. To consider both similarity degree and news title semantics simultaneously when tracking user-interested news events, the modified cosine measure is expressed as

$$Msim(A, B) = \left(1 - \frac{H_{AB}}{Max(H)}\right) \times Sim(A, B) \tag{6}$$

where $Max(H)$ represents the maximum hamming distance among the user-interested news event A with all tracking news events.

In the second phase, the modified cosine measure is further employed to evaluate similarities between the user-interested news event A and candidate news stories/categories obtained during the first phase. In the second phase, when similarities between the user-interested news event A and news events categorized in the candidate stories/categories measured by the modified cosine measure exceed an assigned threshold, then the news events in the candidate news stories/categories are viewed as qualified news events. Only qualified news events can participate in computing the average modified similarity for judging user-interested news events. That is because, although the classification accuracy rate of Google news is very high, the employed automated process cannot absolutely guarantee no any misclassified news events in some news category. To consider qualified news events aims at filtering out misclassified news events while calculating the average modified similarity to recommend news stories/category with highest average modified similarity to individual users. This consideration is helpful to promote the accuracy rate of tracking user-interested news events as well as reduce computational complexity. The proposed average modified similarity is formulated as

$$Avg\_Msim(A, C_j) = \frac{\sum_{l=1}^{n} Msim(A, N_{l(q)})}{n}, \quad j = 1, 2, \ldots, p \tag{7}$$

where $Avg\_Msim(A, C_j)$ represents the average modified similarity of the user-interested news event A with the set of the qualified news events contained in the jth candidate story/category, $N_{l(q)}$ is the lth qualified new event categorized in the jth candidate news category C_j, p is the total number of candidate news stories/categories obtained in the first phase, and n is the number of qualified news events in the jth candidate news stories/category.

5.3 The operation procedure of the proposed two-phase scheme for tracking user-interested news events

Finally, the proposed two-phase scheme employs the average modified similarity to recommend news stories/category with highest average modified similarity to individual users. In summary, the proposed two-phase scheme for tracking user-interested news events can be summarized as

Step 1. In the first phase, compute the similarities between user-interested news events and all news stories/categories contained in the news database by (3).

Step 2. When the similarities between a user-interested news event and the news stories/categories computed in **Step 1** exceed an assigned similarity threshold, then these news stories/categories are retained as candidate news stories/categories, and the proposed scheme enters the second phase, i.e., **Step 3**; otherwise, the news stories/categories are filtered out from the consideration of news stories/categories.

Step 3. When the similarities between a user-interested news event and the news events categorized in the candidate news stories/categories—computed in **Step 2**—exceed an assigned similarity, then these news events are viewed as qualified news events and are preserved for computing the average modified similarity, i.e., **Step 4**; otherwise, the news events are eliminated from the candidate stories/categories.

Step 4. Compute the average modified similarities between the user-interested news event and the qualified news events in the candidate news stories/categories obtained in **Step 3** by (7).

Step 5. Rank the average modified similarities computed in **Step 4** and recommend the news story/category with the highest average modified similarity as the user-interested news story/category to individual users.

6 Experiments

The personalized e-news event monitoring agent system was published on the web site http://dlll.nccu.edu.tw/~rank/navs/index.php to provide e-news monitoring service for tracking user-interested Chinese news events. Currently, the proposed system was implemented on a high-performance computer server with Linux operation system and two processors. The Apache web server, PHP language, and MySQL database were employed to develop the proposed agent system. To demonstrate the performance of the proposed agent system, the system randomly selected news events from the news database as user-interested news events to perform the evaluation of the recall and precision rates while tracking user-interested news events. Additionally, our experiments also respectively compare the accuracy rates of tracking user-interested news events using the traditional Chinese word segmentation system CKIP and the proposed Chinese word segmentation system ECScanner with new word extension mechanism.

6.1 The database information of the gathered Google news

At present, the news crawler agent has collected a large number of news events from Google news site [38] and stored them in MySQL database. Table 1 illustrates the summarization information of the gathered Google news. The summarization information indicates that there are totally 1191065 news events, which are contained in 60269 categories, downloaded from Google news site.

Additionally, this study found that the number of news events per category is between 2 to 20 in most parts of news categories from the statistic results listed in Table 2. In other words, about eighty percent news categories only contain few number of news events. This phenomenon

Table 1 The summarization information of the gathered Google news

Item	Description
The dates of collection news	From 2004-11-9 to 2006-6-16; about 584 days
The number of news events	1191065 news events
The number of news categories	60269 categories
The average number of news events per day	2039.49 news events / day
The average number of news events per category	19.76 news events / category
The news category with maximum number of news events	3007 news events
The news category with minimum number of news events	2 news events

Table 2 The statistic information of Google news based on the number of news events per category

The range of the number of news events	The number of news categories	Percentage of news categories
2–5	13735	22.79%
5–10	18201	30.20%
10–20	14684	24.36%
20–50	9521	15.80%
50–100	2559	4.24%
100–200	1023	1.70%
200–500	424	0.70%
500–1000	95	0.16%
1000–2000	23	0.04%
2000–	4	0.01%

leads to difficulty while determining the statistic parameters of document-frequency-based new word discovery scheme [39]; therefore, this study presents a uniformity-based new words discovery scheme mentioned in Sect. 4.2 to overcome this problem.

Moreover, to evaluate the performance of tracking user-interested news events, the total number of news categories sampled from the entire gathered news database is 1436, the total number of news events contained in the sampled news categories is 30071, and the number of news events randomly selected from the gathered news database as user-interested news events for evaluating the performance of tracking news events is 1000. Table 3 shows the detailed information of the sampling news events in the experiment.

In our experiments, the proposed ECScanner with new words' extension mechanism and the Chinese word segmentation system CKIP, were simultaneously applied to perform Chinese word segmentation process in order to compare the performance of tracking user-interested news events. The initial word lexicon in the proposed Chinese word segmentation system ECScanner which totally contains 50104 (i.e. $39535 + 10569 = 50104$) Chinese words was obtained mainly from the word lexicon in Academia Sinica [40] and CKIP with unknown term guessing scheme [19]. Furthermore, the total number of new words discovered from the archival news database by the proposed new words' extension approach for ECScanner is 1960 (i.e. $2652 - 692 = 1960$). This is because 692 words from the intersection words between both the lexicon databases discovered by CKIP with unknown term guessing scheme and the proposed uniformity-based new words' extension scheme must be deducted. Therefore, about four percent new words are included into the lexicon of ECScanner. Restated, the total number of Chinese words in the word lexicon of ECScanner is 52064 (i.e. $50104 + 1960 = 52064$) in this experiment. Table 4 shows the distributions of lexicon words from 2-grams to 8-grams based on the sampled 1436 news categories. Table 5 displays the comparison of new words discovered by CKIP with unknown term guessing scheme and the proposed uniformity-based words' extension scheme.

6.2 Tracking user-interested news events based on Chinese word segmentation system CKIP using single-phase scheme

Next, the proposed cosine measure with an assigned threshold as (3) was used to measure the similarity for tracking user-interested news events based on Chinese word segmentation system CKIP. In this approach, the union terms' set of a user-interested news event with a news event contained in a randomly sampling news category is used to construct the

Table 3 The detailed information of the sampling news events in our experiment

Item	Description
The total number of news events	30071 news events
The total number of news categories	1436 categories
The average number of news events per day	1768.88 stories/day
The average number of news events per category	20.94 news events/category
The maximum number of news events in the sampling news categories	708 news events
The minimum number of news events in the sampling news categories	2 news events

Table 4 The distributions of lexicon words from 2-grams to 8-grams based on the sampled 1436 news categories listed in Table 3

The gathered scheme of Chinese terms	2-grams	3-grams	4-grams	5-grams	6-grams	7-grams	8-grams	Total
Chinese terms gathered from Academia Sinica lexicon [40]	27079	8902	3339	159	44	11	1	39535
Chinese terms gathered from CKIP with unknown word guessing scheme based on the sampled 1436 news categories	6156	3080	896	297	92	36	12	10569
Chinese terms gathered from the proposed uniformity-based new words discovery scheme based on the sampled 1436 news categories	677	445	1530	–	–	–	–	2652

Table 5 The comparison of new words discovered by CKIP with unknown word guessing scheme and the proposed uniformity-based new words' extension scheme

Proposed scheme	The intersection words between both the lexicon databases discovered from CKIP with unknown term guessing scheme and the proposed uniformity-based new words' discovery scheme	The difference words between both the lexicon databases discovered from CKIP with unknown term guessing scheme and the proposed uniformity-based new words' discovery scheme
The proposed uniformity-based new words' discovery scheme	692 (26.09%)	1960 (73.91%)

terms' vectors for evaluating the similarity of user-interested news event with sampled news category by cosine measure. If similarities between a user-interested news event with all randomly testing news stories/categories stored in the news database are lower than a similarity threshold determined by the user, then the proposed e-news event monitoring agent system will treat the user-interested news event as unknown news events; otherwise, the proposed system will recommend the tracking news category with the highest similarity among all tracking news categories as a user-interested news event. To demonstrate the performance of the proposed agent system, the system randomly selected news events from some news category archived in the news database as user-interested news events for simulating tracking user-interested news events to perform the evaluation of the recall and precision rates. If the decision category of a randomly selected news event judged by the proposed scheme is the same with its original category categorized by the Google news site, then the outcome of tracking user-interested news event is viewed as correct; otherwise, it is viewed as incorrect. Therefore, the proposed system can automatically evaluate the recall and accuracy rates based on the outcomes of automatically grouping related headlines into a news story/category in Google news site. Table 6 displays the experimental results of tracking user-interested news events using a single cosine measure based on Chinese word segmentation system CKIP.

Fig. 7 reveals the relationships of recall, unknown and precision rates with various thresholds based on Chinese word segmentation system CKIP. From the experimental results illustrated as Fig. 7, the precision rate will be obviously promoted while using a high similarity threshold, but the recall rate will be descended with using a high similarity threshold. Additionally, the experimental results also reveal that the promotion of precision rate derives from the reduction of the recall rate as well as the promotion of unknown

Table 6 The experimental results of tracking user-interested news events using cosine measure with single threshold based on Chinese word segmentation system CKIP

Similarity threshold	The number of correct news events	The number of incorrect news events	The number of unknown news events	Recall rate	Unknown rate	Precision rate
0.1	685	311	4	**68.50%**	**0.40%**	68.78%
0.2	676	291	33	67.60%	3.30%	69.91%
0.3	651	235	114	65.10%	11.40%	73.48%
0.4	575	171	254	57.50%	25.40%	77.08%
0.5	480	108	412	48.00%	41.20%	**81.63%**

Table 7 The experimental results of tracking user-interested news events using cosine measure with single threshold based on Chinese word segmentation system ECScanner

Similarity threshold	The number of correct news events	The number of incorrect news events	The number of unknown news events	Recall rate	Unknown rate	Precision rate
0.1	724	272	4	**72.40%**	**0.40%**	72.69%
0.2	715	248	37	71.50%	3.70%	74.25%
0.3	683	196	121	68.30%	12.10%	77.70%
0.4	596	140	264	59.60%	26.40%	80.98%
0.5	478	80	442	47.80%	44.20%	**85.66%**

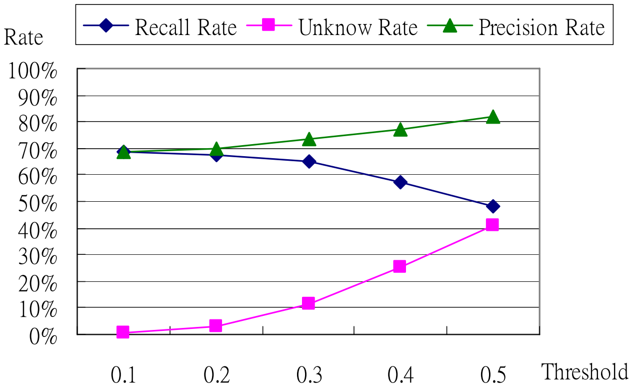


Fig. 7 Relationships of recall, unknown and precision rates with various thresholds based on Chinese word segmentation system CKIP

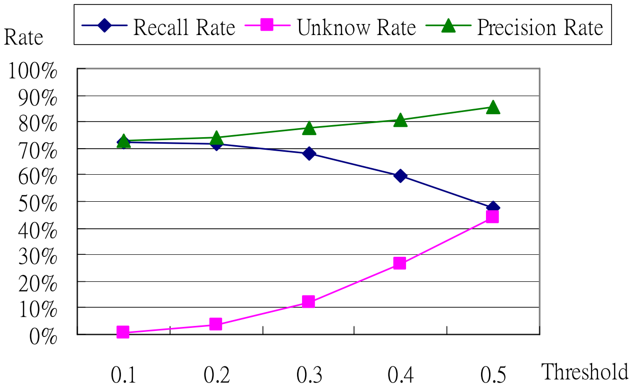


Fig. 8 Relationships of recall, unknown and precision rates with various thresholds based on Chinese word segmentation system ECScanner

rate. It is obviously that this approach is not a good scheme to perform tracking user-interested news events. The proposed average modified similarity measure with dual thresholds will reveal affective promotion while tracking user-interested news events, and called as two-phase approach in this study.

6.3 Tracking user-interested news events based on Chinese word segmentation system ECScanner using single phase scheme

Table 7 displays the experimental results of tracking user-interested news events using cosine measure with a single threshold based on the proposed Chinese word segmentation system ECScanner. Moreover, Fig. 8 reveals the relationships of recall, unknown and precision rates with various thresholds based on Chinese word segmentation system ECScanner. To compare with the experimental results listed in Table 6, the experimental results listed in Table 7 demonstrate that the proposed Chinese word segmentation system ECScanner with new words' extension mechanism is indeed helpful to promote the accuracy rate of tracking user-interested news events.

Table 8 The experimental results of tracking user-interested news events using the modified similarity-measure with dual thresholds based on Chinese word segmentation system CKIP

The first threshold	The second threshold	The number of correct news events	The number of incorrect news events	The number of unknown news events	Recall rate	Unknown rate	Precision rate
0.1	0.4	693	239	68	69.30%	**6.80%**	74.36%
	0.5	703	165	132	**70.30%**	13.20%	80.99%
	0.6	672	133	195	67.20%	19.50%	83.48%
	0.7	656	91	253	65.60%	25.30%	87.82%
	0.8	603	60	337	60.30%	33.70%	**90.95%**
0.2	0.4	682	219	99	**68.20%**	**9.90%**	75.69%
	0.5	681	157	162	68.10%	16.20%	81.26%
	0.6	656	123	221	65.60%	22.10%	84.21%
	0.7	646	84	270	64.60%	27.00%	88.49%
	0.8	592	56	352	59.20%	35.20%	**91.36%**
0.3	0.4	650	199	151	**65.00%**	**15.10%**	76.56%
	0.5	647	147	206	64.70%	20.60%	81.49%
	0.6	627	115	258	62.70%	25.80%	84.50%
	0.7	619	76	305	61.90%	30.50%	89.06%
	0.8	567	50	383	56.70%	38.30%	**91.90%**

6.4 Tracking user-interested news events based on Chinese word segmentation system CKIP using two-phase scheme

To improve the flaw of high recall rate and low precision rate while tracking user-interested news events by the service provided in Google news alert, this study presents a two-phase scheme using the proposed modified similarity measure to track user-interested new events based on Chinese word segmentation system CKIP. Our goal aims at promoting the precision rate as well as also keeping a high recall rate while tracking user-interested news events. From the experimental results mentioned in Sects. 6.2 and 6.3, this study summarized that the recall rate is negative relevant with the similarity threshold of the used cosine measure. To keep a high recall rate, the proposed two-phase scheme employed the cosine measure with a low similarity threshold to filter out irrelevant news categories in the first phase, and then applied the average modified similarity measure with a high similarity threshold in the second phase to further eliminate those irrelevant news events, thus obtaining a high precise rate as well as keeping a satisfied recall rate while tracking user-interested news events.

Table 8 displays the experimental results of tracking user-interested news events using the proposed two-phase scheme based on Chinese word segmentation system CKIP. From the experimental results listed in Table 8, this study found that the recall rate almost keeps unchangeableness under the setting first similarity threshold even the second similarity threshold is increased. Consequently, this study improved the precision rate by promoting the second threshold while tracking user-interested news events.

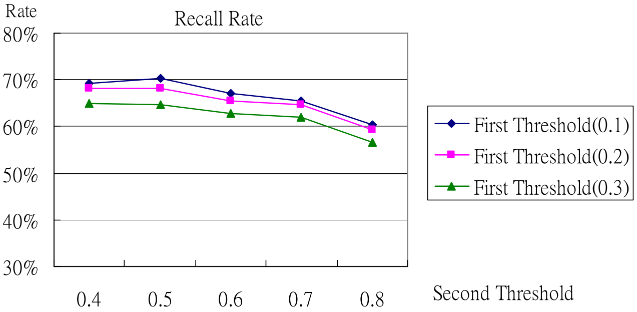


Fig. 9 Relationships of recall rate with various dual thresholds based on Chinese word segmentation system CKIP

Figure 9 shows the relationships of recall rate with various thresholds for the proposed two-phase scheme based on Chinese word segmentation system CKIP. The results also reveal that using a low threshold in the first phase benefits the promotion of recall rate. Figure 10 shows the relationships of precision rate with various thresholds for the proposed two-phase scheme based on Chinese word segmentation system CKIP. The results reveal that using a high threshold in the second phase benefits the promotion of the precision rate regardless of the similarity thresholds set in the first phase. Figure 11 shows the relationships of unknown rate with various thresholds based on Chinese word segmentation system CKIP. The results reveal that using

Table 9 The experimental results of tracking user-interested news events using the modified similarity-measure with dual thresholds based on Chinese word segmentation system ECScanner

The first threshold	The second threshold	The number of correct news events	The number of incorrect news events	The number of unknown news events	Recall rate	Unknown rate	Precision rate
0.1	0.4	702	220	78	70.20%	**7.80%**	76.14%
	0.5	715	140	145	**71.50%**	14.50%	83.63%
	0.6	685	118	197	68.50%	19.70%	85.31%
	0.7	650	88	262	65.00%	26.20%	88.08%
	0.8	595	53	352	59.50%	35.20%	**91.82%**
0.2	0.4	685	202	113	68.50%	**11.30%**	77.23%
	0.5	700	124	176	**70.00%**	17.60%	84.95%
	0.6	668	105	227	66.80%	22.70%	86.42%
	0.7	638	80	282	63.80%	28.20%	88.86%
	0.8	584	48	368	58.40%	36.80%	**92.41%**
0.3	0.4	660	175	165	66.00%	**16.50%**	79.04%
	0.5	667	118	215	**66.70%**	21.50%	84.97%
	0.6	641	99	260	64.10%	26.00%	86.62%
	0.7	613	74	313	61.30%	31.30%	89.23%
	0.8	564	45	391	56.40%	39.10%	**92.61%**

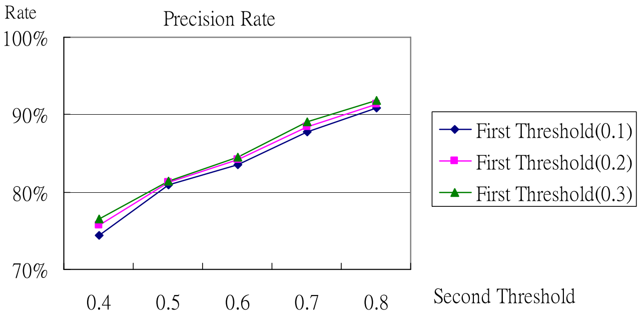


Fig. 10 Relationships of precision rate with various dual thresholds based on Chinese word segmentation system CKIP

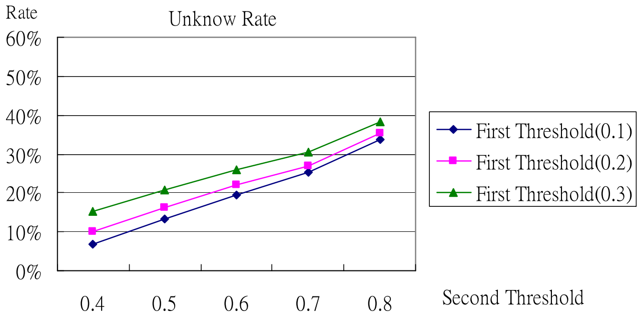


Fig. 11 Relationships of unknown rate with various dual thresholds based on Chinese word segmentation system CKIP

a small threshold in the first phase can obtain a low unknown rate. In summary, the proposed two-phase scheme concluded that using a low similarity threshold in the first phase and a high threshold in the second phase to obtain a high precision rate, an acceptable unknown rate and a satisfied recall rate while tracking user-interested news events.

6.5 Tracking user-interested news events based on Chinese word segmentation system ECScanner using two-phase scheme

Similarly, the proposed ECScanner with new words' extension mechanism also provides benefits to promote the accuracy rate of tracking user-interested news events for the proposed two-phase scheme. Table 9 illustrates the experimental results of tracking user-interested news events using the modified similarity-measure with dual thresholds based on Chinese word segmentation system ECScanner. Figures 12, 13 and 14 show the relationships of the recall rate, precision rate and unknown rate with various thresholds for the proposed two-phase scheme based on Chinese word segmentation system ECScanner, respectively. The experimental results confirmed that the promotion of accuracy rates of tracking user-interested news events is very obvious specially while using the proposed two-phase scheme with low first threshold and high second threshold for tracking user-interested news events.

Table 10 A randomly selected news title for identifying entire sentence as several separated words by linguistic experts, CKIP with unknown term guessing scheme and the ECscanner with the proposed uniformity-based new words discovery scheme

New title randomly selected from the archival news database	綠營:泛藍提罷免或倒閣政治鬥爭意圖奪權 (Pan-green alliance said that pan-blue alliance push the recall vote or topple the cabinet, and the purpose of politics strife is to take over power)
The Chinese word segmentation results by linguistic experts	綠營(Pan-Green Alliance)/泛藍(Pan-Blue Alliance)/罷免(Recall Vote)/倒閣(Topple the Cabinet)/政治(Politics)/鬥爭(Strife)/意圖(Purpose)/***奪權 (Take Over Power)*
The Chinese word segmentation results by CKIP with unknown word guessing scheme	綠營(Pan-Green Alliance)/*泛藍提(Pan-Blue Alliance Push)/罷免(Recall Vote)/倒閣(Topple the Cabinet)/政治(Politics)/鬥爭(Strife)/意圖(Purpose)
The Chinese word segmentation results by the proposed uniformity-based new word discovery scheme	綠營(Pan-Green Alliance)/*泛藍提(Pan-Blue Alliance Push)/罷免(Recall Vote)/倒閣(Topple the Cabinet)/政治(Politics)/鬥爭(Strife)/意圖(Purpose)/***奪權 (Take Over Power)*

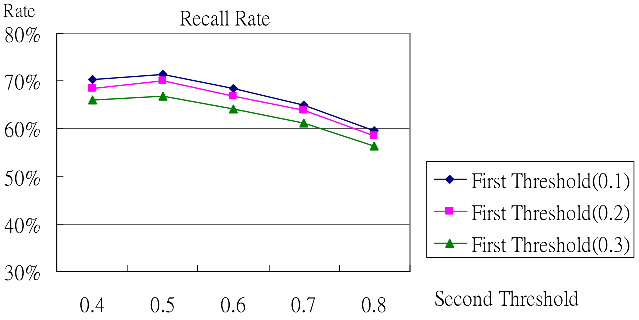


Fig. 12 Relationships of recall rate with various dual thresholds based on Chinese word segmentation system ECScanner

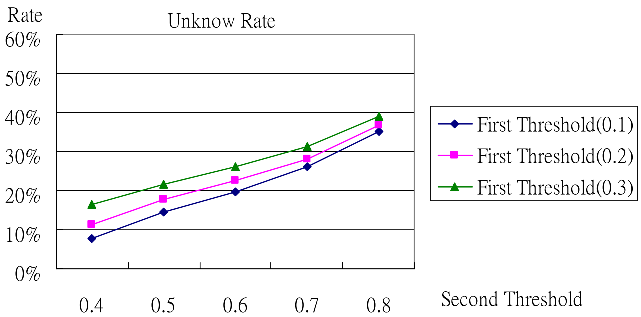


Fig. 14 Relationships of unknown rate with various dual thresholds based on Chinese word segmentation system ECScanner

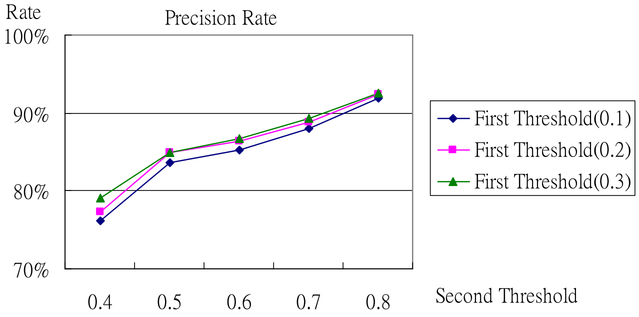


Fig. 13 Relationships of precision rate with various dual thresholds based on Chinese word segmentation system ECScanner

6.6 Evaluating performance for Chinese word segmentation system ECScanner with new words' discovery scheme

In this section, the precision rates of Chinese word segmentation schemes, which include identifying Chinese news titles as separated words by linguistic experts, CKIP with unknown term guessing scheme and the ECScanner with the proposed uniformity-based new word discovery scheme, are compared, respectively. Table 10 displays a randomly selected Chinese news title for identifying entire sentence as several separated words by various schemes. Moreover, Table 11 illustrates the statistic results of Chinese word segmentation by CKIP with unknown term guessing scheme and the ECScanner with the proposed uniformity-based new word discovery scheme for the news title listed in Table 10. In the results, the notation "**" stands for the new words that have been identified by Chinese language experts, but the CKIP with unknown term guessing scheme cannot identify them well. The notation "*" stands for the incorrect new words identified by both the CKIP with unknown term guessing scheme and ECscanner with the proposed scheme of new word discovery. The results show that the ECscanner with the proposed new words' discovery scheme is superior to the CKIP with unknown words guessing scheme in terms of the number of identifying correct words and the number of unknown words. Furthermore, one hundred news titles are randomly sampled from the gathered news database for evaluating the performances of two Chinese word segmentation systems. Table 12 illustrates the compared results for two Chinese word segmentation systems. The experimental results demonstrate that the proposed Chinese word segmentation system ECScanner with new words' extension mechanism is indeed helpful to promote the accuracy and recall rates of Chinese word segmentation as well as reduce the unknown rate of Chinese word segmentation.

Table 11 The statistic results of Chinese word segmentation by CKIP with unknown term guessing scheme and the ECScanner with the proposed uniformity-based new word discovery scheme for the news title listed in Table 10

Compared scheme	The number of correct words	The number of incorrect words	The number of unknown words
The Chinese word segmentation results by CKIP with unknown term guessing scheme	6	1	1
The Chinese word segmentation results by ECScanner with the proposed uniformity-based new word discovery scheme	7	1	0

Table 12 The performance comparison of Chinese word segmentation for CKIP with unknown term guessing scheme and ECScanner with the proposed uniformity-based new word discovery scheme for one hundred news titles randomly selected from the gathered news database

Compared scheme	Recall rate	Precision rate	Unknown rate
The Chinese word segmentation results by CKIP with unknown word guessing scheme	88.96%	95.39%	6.75%
The Chinese word segmentation results by the proposed entropy-based new word discovery scheme	93.25%	96.82%	3.68%

7 Discussion

This section analyzes the characteristics of our proposed scheme and discusses some possible enhancements that should be investigated in the future.

7.1 Determining the uniformity threshold for the proposed uniformity-based new words discovery scheme

Currently, the uniformity threshold of the proposed uniformity-based new words discovery scheme was determined by linguistic experts according to heuristic trying processes for discovering the possible Chinese new words from Google news corpuses. At present, how to appropriately determine the used uniformity threshold lacks a good scheme for the proposed uniformity-based new words discovery scheme. In general, a strict threshold value will reduce the number of possible candidate new words. On the contrary, a loose threshold value will lead to over large number of candidate new words. To avoid incorrect news words included in the lexicon, linguistic experts can determine that the discovered new words are formally accepted as new words or rejected as un-meaningful words by an administrator interface. This mechanism can avoid that incorrect news words are included in the lexicon. From the experimental results of this study show that the developed Chinese words' discovery scheme is indeed helpful to lexicon-based Chinese word segmentation system to promote the accuracy rate of tracking user-interested news events.

7.2 Constructing automatically synonyms for promoting the accuracy rate of tracking user-interested news events

The most search engines like Google alert track user-interested new events utilizing a full-indexing scheme with keyword matching. However, the key word matching scheme usually suppose the user can give an appropriate searching query while using a searching engine for information retrieval, so that using over one searching query occurs frequently. If a user query can be automatically extended based on a synonyms lexicon database, then the extended mechanism can effectively promote the recall rates of search engines. Clearly, the synonyms lexicon database will also provide benefits for the proposed personalized e-news monitoring agent system for promoting the recall rate of tracking user-interested news events.

7.3 How the proposed system could perform with other language news events

Actually, the proposed two-phase approach for tracking user-interested news events can also be employed to track the other language news events except Chinese news events. The main difference is that Chinese news events must perform word segmentation process to identify meaningful separated phrases for tracking user-interested news events, but the identification of distinct words in texts is natural and trivial task for some languages, such as English news event.

7.4 Incorporating how important a term into the similarity measure to reflect different influences of different terms in a news title

Currently, this study employed the binary term weight to represent a news event as the term vector for the used cosine similarity measure while tracking user-interested news events. Therefore, each term that appears in a news title will be treated as identical importance. Although enabling users to specify how important a term may be helpful to promoting precision and accuracy of tracking user-interested news events. However, specifying the term importance of each separated term in a news event to track user-interested news events will lead to degenerating the proposed topic-tracking-based approach into the keyword-based approach if some term is specified an over high term weight by a user. Actually, the proposed topic-tracking-based approach mainly focuses on improving the drawback of the keyword-based scheme used in Google news alert for tracking user-interested news event more accurately. Therefore, it is a logical approach that considering each separated term that appears in a news title has the identical importance because a complete news event consists of all separated terms. No any term in a news title should be particularly ignored or overemphasized its importance while tracking the long-term developing tracks of a news story by the proposed topic-tracking-based approach. Therefore, permitting individual users to specify term importance for each separated term that appears in a news title for tracking user-interested news events was not considered in the study.

7.5 How to identify the difference of two news categories with the identical average modified similarity but containing different amounts of news events

In Google news site, a news category containing a large amount of news events indicates that the descriptive news event is concerned by much more news media. That is, this news can be viewed as a hot, focus or popular news event. Therefore, a news category containing large amount of news events is more important than a news category containing small amount of news events. Based on the aspect, this study did not definitely conclude that a news category with much more amount of news events than another news category is more relevant with the user-interested news event when these two news categories that contain different amounts of news events have the identical average modified similarity. To face this problem, the proposed scheme of tracking user-interested news events will simultaneously recommend the news categories, which have identical average modified similarity with the user-interested news event, to users even they contain different amounts of news events. To do so, users will not lose any interested news information, but the news category with much more news events than another news category should indeed be assigned a higher priority to be recommend to users because it is a relatively more important news event. This mechanism can be easily implemented by ranking the recommending order based on the amount of news events that contain in each news category.

8 Conclusion

This study presents a personalized e-news event monitoring agent system based on the proposed two-phase scheme that retrieves user-interested news events from the Google News site and dispatches immediately the discovered messages to individual users using short messages to cell phones or by e-mail to mailbox. In real-world applications, the proposed agent system save a lot of manpower in monitoring user-interested news events occurring worldwide and shorten search time for everyday news events. The proposed agent system accelerates the valuation of news events because the news events often have time validity. Results of this study demonstrate that the proposed agent system—based on setting appropriate parameters and performing a Chinese word segmentation process using the proposed ECScanner with a new word extension mechanism for the proposed two-phase scheme—obtains a satisfactory accuracy rate while tracking user-interested Chinese news events.

References

1. Cheung P-S, Huang R, Lam W (2004) Financial activity mining from online multilingual news. In: The international conference on information technology: coding and computing
2. Fung GPC, Yu JX, Lam W (2003) Stock prediction: integrating text mining approach using real-time news. In: IEEE international conference on computational intelligence for financial engineering, pp 395–402
3. Mittermayer M-A (2004) Forecasting intraday stock price trends with text mining techniques. In: The 37th Hawaii international conference on system sciences, pp 1–10
4. Wiithrich B, Permunetilleke D, Leung S, Cho V, Zhang J, Lam W (1998) Daily prediction of major stock indices from textual www data. In: Proceedings of the 4th international conference on knowledge discovery and data mining, KDD-98
5. Fawcett T, Provost F (1999) Activity monitoring: noticing interesting changes in behavior. In: Chaudhuri, Madigan (eds) Proceedings on the fifth ACM SIGKDD international conference on knowledge discovery and data mining, San Diego, CA, pp 53–62
6. Wuthrich B et al (1998) Daily stock market forecast from textual web data. In: IEEE International conference on systems, man, and cybernetics, pp 1–6
7. Peramunetilleke D, Wong RK (2002) Currency exchange rate forecasting from news headlines. In: Proceedings of the thirteenth Australasian database conference
8. Nesbitt KV, Barrass S (2004) Finding trading patterns in stock market data. IEEE Comput Graph Appl 24(5):45–55

9. Kuo RJ, Chen CH, Hwang YC (2001) An intelligent stock trading decision support system through integration of genetic algorithm based fuzzy neural network and artificial neural network. Fuzzy Sets Syst 118(1):21–45
10. Shan NA, Elbahesh EM (2004) Topic-based clustering of news articles. In: Proceedings of the 42th annual southeast regional conference, pp 412–413
11. Maria N, Silva MJ (2000) Theme-based retrieval of web news. In: SIGIR, July 2000, pp 354–356
12. Kurtz AJ, Mostafa J (2003) Topic detection and interest tracking in a dynamic online news source. In: Proceedings of the 2003 joint conference on digital libraries
13. Lam W, Cheung P-S, Huang R (2004) Mining events and new name translations from online daily news. In: Proceedings of the 4th ACM/IEEE-CS joint conference on digital libraries, pp 287–295
14. Allan J, Papka R, Lavrenko V (1998) On-line new event detection and tracking. In: SIGIR, pp 37–45
15. Lee C-S, Jian Z-W, Huang L-K (2005) A fuzzy ontology and its application to news summarization. IEEE Trans Syst Man Cybern Part B: Cybern 35(5):859–880
16. Michael JAB, Gordon L (2004) Data mining techniques for marketing, sales, and customer relationship management. Indianapolis, Wiley
17. Google alerts. Web available at http://www.google.com/press/descriptions.html#alerts
18. Foo S, Li H (2004) Chinese word segmentation and its effect on information retrieval. Inf Process Manag 40:161–190
19. Chinese knowledge information processing (CKIP). Web available at http://140.109.19.112/
20. Ma W-Y, Chen K-J (2003) Introduction to CKIP Chinese word segmentation system for the first international Chinese word segmentation bakeoff. In: Proceedings of ACL, second SIGHAN workshop on Chinese language processing, pp 168–171
21. ECScanner (An Extension Chinese Lexicon Scanner). Web available at http://dlll.nccu.edu.tw/~rank/ecscanner/
22. Google news. Web available from: http://www.google.com/press/descriptions.html#news
23. Google advanced search. Web available at http://www.google.com/press/descriptions.html#special
24. Caglayan A, Harrison C (1997) Agent sourcebook: a practical guide to introducing agent technology into your business applications. New York, Wiley
25. Yeh CL, Lee HJ (1991) Rule-based word identification for mandarin Chinese sentences—a unification approach. Comput Process Chin Oriental Lang 5:97–118
26. Zhang M-Y, Lu Z-D, Zou C-Y (2004) A Chinese word segmentation based on language situation in processing ambiguous words. Inf Sci 162(3–4):275–285
27. Chen KJ, Liu SH (1992) Word identification for mandarin Chinese sentences. In: Proceedings of COLING, pp 101–107
28. Dee HM (1985) Introduction to natural language processing. Va.Reston, Reston
29. Huang CR, Chen KJ, Chang LL (1997) Segmentation standard for Chinese natural language processing. Int J Comput Linguist Chin Lang Process 2(2):47–62
30. He S, Zhu J (2000) A bootstrap method for Chinese new words extraction. IEEE Int Conf Acoust Speech, Signal Process 1(7–11):581–584
31. Nie JY, Brisebois M, Ren XB (1996) On Chinese text retrieval. In: Proceedings of SIGIR'96, pp 225–233
32. Wu ZM, Tseng G (1993) Chinese text segmentation for text retrieval: achievements and problems. J Am Soc Inf Sci 44(9):532–542
33. Wu ZM, Tseng G (1995) ACTS: an automatic Chinese text segmentation system for full text retrieval. J Am Soc Inf Sci 46(2):83–96
34. Chowdhury GG (2004) Introduction to modern information retrieval Facet, London
35. CScanner (A Chinese Lexicon Scanner). Web available at http://technology.chtsai.org/cscanner/
36. Department of Chinese Literature of National Chengchi University. Web available at http://www.chinese.nccu.edu.tw/english/english06/index.htm
37. Dhillon IS, Modha DS (2001) Concept decompositions for large sparse text data using clustering. Mach Learn 42(1):143–175
38. Taiwan version of Google news. Web available at http://news.google.com.tw/
39. Chen KJ, Ma WY (2002) Unknown word extraction for Chinese documents. In: Proceedings of COLING, pp 169–175
40. Chinese word lexicon. Web available at http://www.aclclp.org.tw/use_rlssd_c.php