Expert Systems with Applications 37 (2010) 6651–6662

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Intelligent location-based mobile news service system with automatic news summarization

Chih-Ming Chen *

Graduate Institute of Library, Information and Archival Studies, National Chengchi University, No. 64, Sec. 2, ZhiNan Rd., Wenshan District, Taipei City, 116 Taiwan, ROC

ARTICLE INFO

Keywords:
News summarization
Information retrieval
Mobile device
Location-awareness service
Fuzzy synthetic evaluation

ABSTRACT

The accelerated rapid development of wireless network and mobile computing technologies has increased the convenience of mobile information services for obtaining useful information in our daily lives. The technology has thus become an essential research issue. In particular, location-based services (LBSs) on mobile devices can convey location related information to individual users, thus helping users to obtain helpful and adaptive information. Accordingly, this study presents an intelligent location-based service system to provide local news with summaries to personal handheld PDA (Personal Digital Assistant) device based on the user's location. First, this study proposes a novel multi-document news summarization scheme based on fuzzy synthetic evaluation. This scheme identifies a key paragraph in a news story as a summary based on three evaluation scores, in order to dispatch news information to mobile devices with small screens. Furthermore, to sense a user's location precisely, the study also develops a practicable location-awareness scheme based on GPS (Global Positioning System) signals to identify the position of a user in the Taiwan area. Finally, an intelligent location-based mobile news service system with automatic news summarization for location-based news services was implemented by combining the proposed multi-document news summarization and user location schemes. The proportion of satisfactory news summaries is up to 86%, and the accuracy of the user location-awareness scheme is up to 90%, according to experimental results in multi-document summarization and user location identification. These experimental results show that the proposed system can be successfully applied in real-world applications for personalized location-based news services.

1. Introduction

Numerous paper-based newspapers have been transformed into digital format and published on the Internet, so that digital newspapers are gradually becoming popular electronic media for conveying information immediately. News documents frequently contain rich information that is useful and valuable in decision-making (Cheung, Huang, & Lam, 2004; Fung, Yu, & Lam, 2003). However, immediately obtaining key news information required by an individual user is a fundamental issue. 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. Moreover, mobile phones and Personal Digital Assistant (PDA) increasingly enable people to access the Internet at any time and place, making location-based services increasingly important in our everyday lives. Thus, providing immediate news information associated with user's location significantly benefits individuals in making decisions. For instance, local security news can enhance personal safety while traveling; local traffic news can help a person to plan appropriate travel routes, and local accident news can avoid suffering from accident damage. Hence, an intelligent system providing location-based news services is valuable for many people.

A location-based service (LBS) is an information service that can be accessed by mobile devices through a mobile network, and which exploit the location of the mobile device (Virrantaus, Markkula, Garmash, & Terziyan, 2001). Location-based services for mobile devices include a location-aware collaboration system named NAMBA for shopping and meeting (Yoshino, Muta, & Munemori, 2002); a location-aware medical information system for determining a hospital worker's current location from a hospital information system (Rodriguez, Favela, Martinez, & Munoz, 2004), and a ubiquitous GPS and web enabled mobile search mechanism for retrieving personalized and locally targeted search results (Choi, 2007). Other location-based services for mobile devices, such as location-based games on cellular phones, and context-aware ubiquitous learning system based on learner location for supporting effectively English vocabulary learning, are also already operational LBS applications in the entertainment and education fields (Chen, Li, & Chen, 2007; Rashid, Mullins, Coulton, & Edwards, 2006). Although no intelligent location-based news ser-

* Tel.: +886 2 29393091x88024; fax: +886 2 29384704.
E-mail address: chencm@nccu.edu.tw

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.03.055

vice has been developed for mobile devices based on our best knowledge and literature survey, access to information through mobile phones and other handheld devices is growing significantly, supporting decision-making away from a computer (Otterbacher, Radev, & Kareem, 2008; Yang & Wang, 2007). Large documents are impossible to load and visualize on handheld mobile devices due to the limited screen size, low network bandwidth and low memory capacity of these devices (Yang & Wang, 2007). Therefore, summarization techniques have been adopted to deliver information on handheld devices (Otterbacher et al., 2008; Yang & Wang, 2007).

News summarization is a document summarization method that extracts significant news information for readers (Chen, Kuo, Huang, Lin, & Wung, 2003). The main purpose of a summary is to present the main ideas in a document in less space. The current state of the art of summarization includes single-document summarization through extraction, the beginnings of abstractive approaches to single-document summarization, and a variety of approaches to multi-document summarization (Radev, Hovy, & McKeown, 2002). Identifying the informative paragraphs from single or multiple documents is the main challenge in summarization (Radev et al., 2002). The three major problems on multi-document summarization are recognizing and coping with redundancy, identifying important differences among documents, and ensuring summary coherence even in material from different source documents (Radev et al., 2002). Measuring the quality of a summary has certainly proven to be a difficult problem, principally because no obvious "ideal" summary exists, even for relatively straightforward news articles (Radev et al., 2002).

Therefore, this study aims at developing an intelligent location-based mobile news service system with automatic news summarization for mobile device users. Providing location-based news service to a mobile device involves considering two key issues: identifying the most representative paragraph as news summarization for mobile devices from categorized Google news corpora (Google news, 2009), and developing an accurate user location detection scheme with a low computational load to support the proposed location-based news services on mobile devices. To reach these goals, this study presents a novel fuzzy synthetic evaluation based multi-document news summarization scheme based on categorized Google news corpora, which is suitable for information delivery based on user location to small, mobile devices. Additionally, this study proposes an effective user location detection scheme with low computational load, in which the user is located through GPS signals, to support the proposed locations-based news services. Meanwhile, the proposed schemes have been successfully implemented as an intelligent location-based news service system for delivering news information to handheld mobile devices. Experimental results reveal that the proposed multi-document news summarization scheme can obtain a satisfied news summarization quality for users to read, while the proposed user location-awareness scheme can accurately identify user location to support the proposed location-based news service.

The remainder of this paper is organized as follows. Section 2 presents the system architecture of the proposed intelligent location-based mobile news service system with automatic news summarization. Section 3 describes the proposed multi-document news summarization scheme based on categorized Google news corpora. Section 4 presents a user location-awareness scheme based on GPS signals to support the proposed location-based news services. Section 5 presents our experimental results. Finally, conclusions are stated in Section 6.

2. System design

This section describes in detail the system architecture of the proposed intelligent location-based news service system, and the system components.

2.1. System architecture

This study presents an intelligent location-based mobile news service system containing both the server and client-side subsystems to provide location-based news service. Figs. 1 and 2 show the system architectures of the server and client subsystems, respectively. In the server side subsystem, the news crawler agent extracts news documents from the Google news site, and identifies news metadata, including news title, published medium, reporter, location, date, news body and hyperlink to the original published medium, to the e-news and information database based on regular expression matching. Chinese word segmentation is performed on news titles and news bodies using ECScanner (2009) to perform multi-document news summarization. The Chinese word segmentation process filters out all 1-gram words, because these are non-meaningful in judging informative paragraphs for news summarization. Moreover, the multi-document news summarization agent primarily identifies a single most representative paragraph

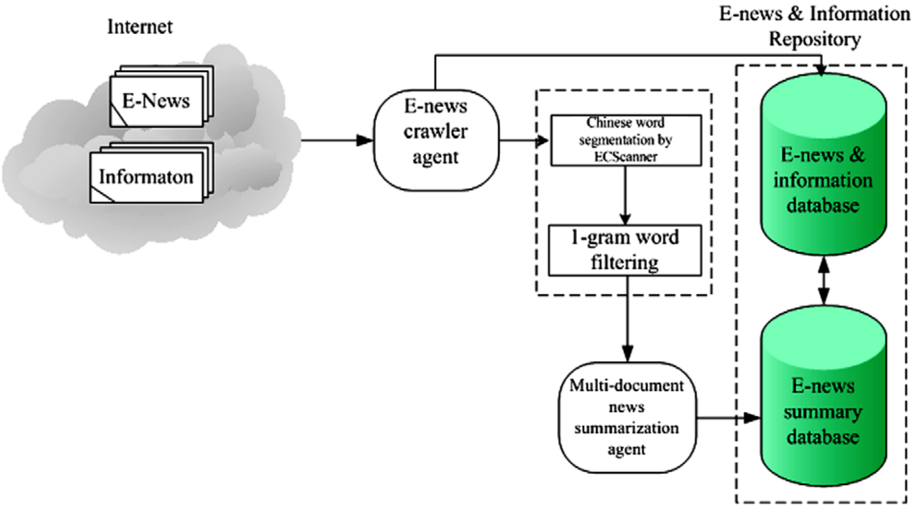

Internet
E-News
Informaton
E-news crawler agent
Chinese word segmentation by ECScanner
1-gram word filtering
Multi-document news summarization agent
E-news & Information Repository
E-news & information database
E-news summary database

Fig. 1. The system architecture of the server side of the proposed intelligent location-based news service system.

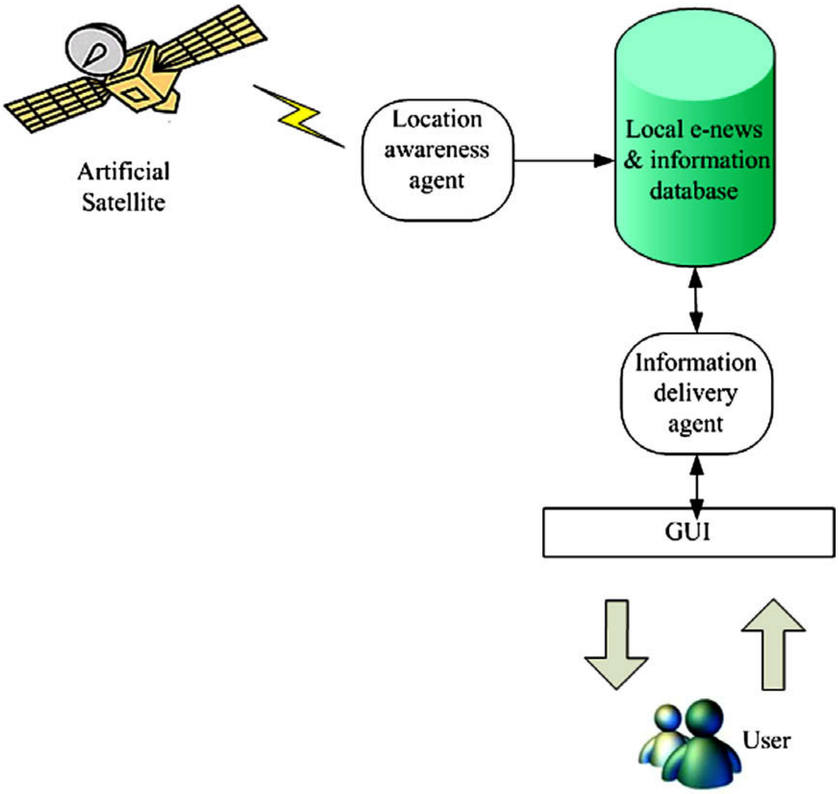

Artificial Satellite
Location awareness agent
Local e-news & information database
Information delivery agent
GUI
User

Fig. 2. The system architecture of the client-side of the proposed intelligent location-based news service system.

as a news summary based on the proposed multi-document news summarization scheme inferred by fuzzy synthetic evaluation, using the three evaluation scores. Finally, all news summaries identified for each Google news story are stored in the e-news summary database to provide a location-based news service for mobile devices. In the client-side subsystem, the location-awareness agent detects user location based on GPS signals, and it assists the information delivery agent to deliver appropriate news to the user PDA mobile device based on the user location. This study employed the proposed average *k*-NN classifier discussed in Section 4.2 to identify user location by GPS signals in the Taiwan area.

2.2. System components

This section introduces and describes the system components of the proposed intelligent location-based mobile news service system.

2.2.1. E-news crawler agent

The e-news crawler agent automatically retrieves news events from the Google news site at the scheduled running time. Metadata are extracted from the extracted news events to parse important metadata, namely the news title, reporter, medium, date, location, body text and hyperlink, to the original published medium. The metadata extraction process extracts important news metadata based on the news templates in the Google news site. Since the news templates in the Google news site are organized using HTML tags, the metadata extraction process identifies news metadata based on the properties of the front and rear HTML tags with extracted metadata. However, the characteristics of HTML tags for extracting news metadata must be identified manually. A software program is then designed to identify useful metadata from HTML documents using regular expressions.

2.2.2. Chinese word segmentation system ECScanner

The Chinese word segmentation process segments a Chinese news title and body into several separate Chinese words. This process influences the summarization quality in identifying the most representative paragraph. Since news events often contain words that seldom appear in previous news events, a Chinese word segmentation system with a fixed word lexicon cannot effectively segment Chinese news titles and bodies. This study extended ECScanner with a new word extension mechanism proposed in our previous study (Chen & Liu, 2009).

2.2.3. Multi-document news summarization agent

The multi-document news summarization agent identifies the most representative paragraph from news articles with different news titles grouped in the same news story. Fuzzy synthetic evaluation is adopted to identify the most representative paragraph based on three evaluation scores, namely length, similarity and information.

2.2.4. E-news and information database

The e-news and information database stores Google news corpora retrieved from the Google news site. The stored information retrieved by the e-news crawler agent with metadata extraction is composed of the news title, reporter, medium, date, location, body text and hyperlink to the original published medium. The location where a piece of news was reported is the most important information for the proposed location-based news services.

2.2.5. E-news summary database

The e-news summary database stories news summaries generated by the multi-document news summarization agent.

The system components of the client-side implemented in user PDA mobile device are next described in detail as follows:

2.2.5.1. Location-awareness agent. The location-awareness agent detects the location of the user, and transmits this to the information delivery agent. In this study, the proposed average *k*-NN classifier was employed to identify the user's location in the Taiwan area based on GPS signals.

2.2.5.2. Local e-news and information database. The local e-news and information database can temporarily store news information from the server side subsystem in the local database system of a PDA mobile device for news services.

2.2.5.3. Information delivery agent. The information delivery agent retrieves the appropriate local news summaries from the e-news and information database, according to the user location detected by the user location-awareness agent, and delivers them into the user's PDA mobile device.

3. Proposed multi-document news summarization scheme based on Google news corpora

This section describes the proposed multi-document news summarization scheme based on Google news corpora in terms of the three evaluation scores. This scheme identifies the most representative paragraph from a Google news story that groups related news articles together.

3.1. Google news with automatic news event grouping

The Google news site currently accesses nearly 10,000 news sources on the Internet by an automatic crawler program. The content from these sources is presented as news stories/categories in a searchable format on the web (Google news, 2009). Leading stories selected automatically by a computer algorithm are treated as headlines on the Google news home page, without regard to political viewpoint or ideology. Google news adopts an automated process to group related news events together into a news story/category, which in some cases enables people to access different viewpoints on the same story/category (Google news, 2009). Evaluation results in this study indicate that the automated process achieves a very high accuracy rate, even it does not lose in manual classification. The Google news service is currently tailored to 22 international audiences (Google news, 2009), and thus provides very appropriate corpora to support the proposed multi-document news summarization scheme.

3.2. Proposed scheme for multi-document news summarization

3.2.1. Flowchart of the proposed multi-document news summarization scheme

Fig. 3 shows the operation flowchart of the proposed multi-document news summarization scheme. The proposed scheme first determines the three evaluation scores, namely length, similarity and information of a paragraph to identify a most representative paragraph in a news story. Restated, the system computes the representativeness of a paragraph in a news story from these three evaluation scores simultaneously. The scores are normalized into the interval [0,1] by taking the corresponding three highest scores to be 1. The news summaries are then identified from the scores using fuzzy synthetic evaluation (Chang, Chen, & Ning, 2001; Kuo & Chen, 2006). This process involves converting the crisp evaluation scores into fuzzy degrees. This study applied the *K*-means clustering algorithm (Kanungo et al., 2002) to determine the centers of the triangle fuzzy membership functions automatically according to the data distribution of each evaluation score. To simplify the fuzzy inference procedures, the number of clusters in the *K*-means clustering algorithm was set to 3. Thus, each evaluation score contained three fuzzy linguistic terms, i.e. low, moderate and high, to fuzzify the normalized evaluation scores. Fuzzy synthetic evaluation method is then adopted to infer the representative degrees of all paragraphs in a news story based on three adjustable weights heuristically determined by language experts.

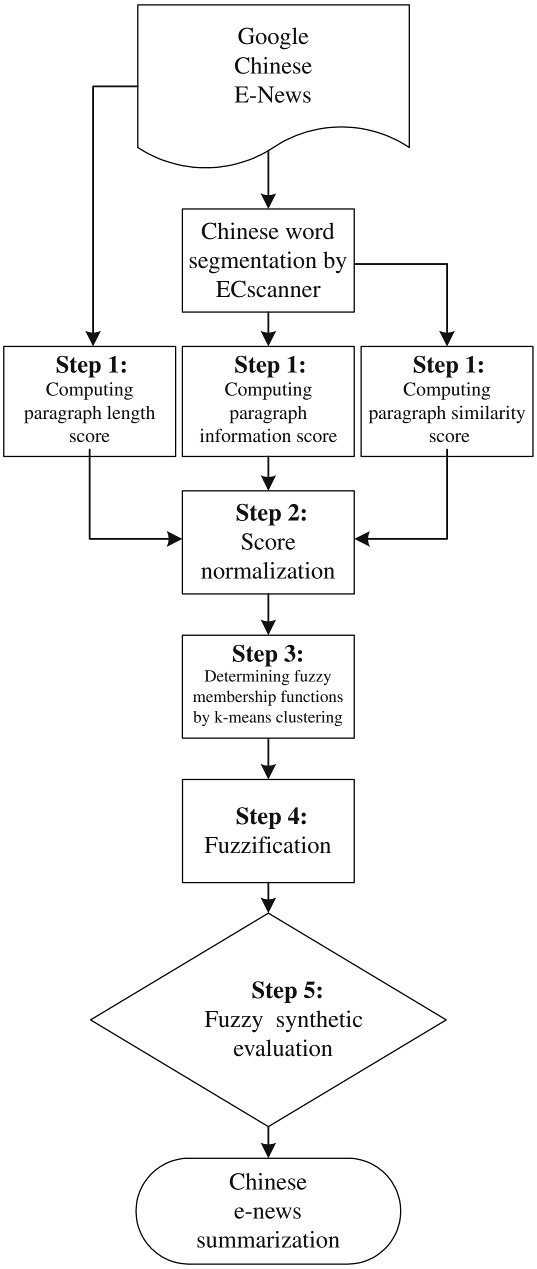

Google
Chinese
E-News
Chinese word segmentation by ECscanner
Step 1: Computing paragraph length score
Step 1: Computing paragraph information score
Step 1: Computing paragraph similarity score
Step 2: Score normalization
Step 3: Determining fuzzy membership functions by k-means clustering
Step 4: Fuzzification
Step 5: Fuzzy synthetic evaluation
Chinese e-news summarization

Fig. 3. The flowchart of the proposed news summarization scheme.

3.2.2. Proposed three evaluation scores for identifying the representative degree of a paragraph in a news story

The representativeness of a paragraph as a summary in a news story is measured in terms of three evaluation scores, namely the length, similarity and information scores. These scores are described as follows.

3.2.2.1. Length score (LS). The average length of a Reuters news summary is 85–90 words according to statistical analysis (Goldstein, Kantrowitz, Mittal, & Carbonell, 1999). The length of a news summary is an important issue in summarizing news articles. Therefore, this study used the length of a paragraph as an indicator of its representativeness in a news story. However, no studies have considered the optimal length of a news summary for readability. This study proposes considering the average length of all news paragraphs in a news story as an appropriate length for a news summary. An excessively short news summary could lose some supplementary news context information, although it still retrieves the key part of a news article. Conversely, an excessively long news summary often contains unimportant or redundant news information, thus reducing readability and reading efficiency. The average length of all news paragraphs in a news story can be calculated as follows:

$$\mathrm{APL} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\mathrm{length}(p_{ij})}{M \times N} \tag{1}$$

where APL denotes the average paragraph length of all news articles in a news story; M indicates the total number of news articles in a news story; N represents the total number of paragraphs across all news articles in a news story, and length (p_{ij}) is the length of the jth paragraph of the ith news article in a news story.

Therefore, the length score of a paragraph in a news story can be computed as

$$\begin{cases} \mathrm{LS}(p_{ij}) = \frac{\mathrm{length}(p_{ij}) - \mathrm{length}_{\max}(P)}{|\mathrm{APL} - \mathrm{length}_{\max}(P)|}; \mathrm{length}(p_{ij}) > \mathrm{APL} \\ \mathrm{LS}(p_{ij}) = \frac{2\times\mathrm{APL} - \mathrm{length}_{\max}(P) - \mathrm{length}(p_{ij})}{|\mathrm{APL} - \mathrm{length}_{\max}(P)|}; \mathrm{length}(p_{ij}) \le \mathrm{APL} \end{cases} \tag{2}$$

where $\mathrm{length}_{\max}$ (P) is the length of the longest news paragraph in a news story.

3.2.2.2. Similarity score (SS). Salton and McGill (1983) and Salton (1989) concluded that a paragraph that is strongly associated with the other paragraphs in an article is highly representative of the subject of this article. Therefore, the representativeness level of a paragraph in a news story is measured using a similarity score. First, the similarity between two paragraphs can be calculated as

$$\mathrm{Sim}(p_{ij}, p_{mn}) = \frac{w(p_{ij}) \cap w(p_{mn})}{\sqrt{n(p_{ij}) \times n(p_{mn})}} \tag{3}$$

where p_{ij} denotes the jth paragraph of the ith news article in a news story; p_{mn} indicates the nth paragraph of the mth news article in a news story; $w(p_{ij}) \cap w(p_{mn})$ is the number of words common to the jth paragraph of the ith news article and the nth paragraph of the mth news article in the same news story; $n(p_{ij})$ is the total number of words in the jth paragraph of the ith news article, and $n(p_{mn})$ is the total number of words in the nth paragraph of the mth news article.

Next, a threshold is defined to determine the points for calculating the similarity score, and formulated as

$$P_{\text{Get-Point}}(p_{ij}, p_{mn}) = \begin{cases} 1, \mathrm{Sim}(p_{ij}, p_{mn}) \ge \theta \\ 0, \mathrm{Sim}(p_{ij}, p_{mn} < \theta \end{cases} \tag{4}$$

where θ denotes a threshold heuristically determined by language experts.

Finally, the similarity score of a paragraph can be computed as

$$\mathrm{SS}(p_{ij}) = \sum_{i=1,m=1}^{M} \sum_{j=1,n=1,j\neq n}^{N} P_{\text{Get-Point}}(p_{ij}, p_{mn}) \tag{5}$$

where M denotes the total number of news articles in a news story, and N indicates the total number of paragraphs of all news articles in a news story.

3.2.2.3. Information score (IS). Generally speaking, humans can approximately guess the news information conveyed in a news article based on the news headline, even without reading the news body. Although headlines reporting the same news event may be different, they generally contain the same key words it. A similar phenomenon occurs in the paragraphs comprising news articles on the same story. Therefore, the words that frequently appear in titles, most news articles and paragraphs in a Google news story are more representative than other words of the story. Therefore, the proposed multi-document news summarization scheme includes an information score to measure the representativeness of a paragraph based on these important words. The information score of the jth paragraph of the ith news article in a news story can be formulated as follows:

$$\begin{aligned} \mathrm{IS}(p_{ij}) = {} & \alpha \times \sum_{k=1}^{n}(\mathrm{TF}(w_{i,j,k}) \times n_{\mathrm{Tk}}) + \beta \times \sum_{k=1}^{n}(\mathrm{DF}(w_{i,j,k}) \times n_{\mathrm{Dk}}) \\ & + \gamma \times \sum_{k=1}^{n}(\mathrm{PF}(w_{i,j,k}) \times n_{\mathrm{Pk}}) \end{aligned} \tag{6}$$

where $w_{i,j,k}$ denotes the kth word of the jth paragraph of the ith news article; $\mathrm{TF}(w_{i,j,k})$ indicates the term frequency of the kth word of the jth paragraph of the ith news article, $\mathrm{DF}(w_{i,j,k})$ is the document frequency of the kth word of the jth paragraph of the ith news article; $\mathrm{PF}(w_{i,j,k})$ is the paragraph frequency of the kth word of the jth paragraph of the ith news article; n_{Tk} is the number frequency in all news titles of the kth word of the jth paragraph of the ith news article; n_{Dk} is the number news articles containing the kth word of the jth paragraph of the ith news article; n_{Pk} is the number of news paragraphs containing the kth word of the jth paragraph of the ith news article; n is total number of words in the jth paragraph of the ith news article in a news story, and α, β, and γ $(\alpha + \beta + \gamma = 1)$ are adjustable weights for $\mathrm{TF}(w_k)$, $\mathrm{DF}(w_k)$, and $\mathrm{PF}(w_k)$, respectively.

3.2.3. Fuzzy synthetic evaluation method for identifying the representativeness of a paragraph in a news story based on three considered evaluation scores for news summarization

This section describes the proposed multi-document news summarization scheme, which simultaneously considers three indicators, namely length, similarity and information, to assess the important levels of a paragraph in a Google news story based on fuzzy synthetic evaluation. Fuzzy synthetic evaluation produces results that are more scientific and reasonable than the normal methods, because it addresses the fuzziness of the gradual change from low to high level of the quality of the affected news summarization. Every main factor affecting the news summarization quality needs to be determined, together with setting evaluating indicators, evaluating set and membership function. The weight of every indicator, and the degree of membership, are then determined to obtain the synthetic degree of membership, from which the news summarization quality is obtained. Before performing fuzzy synthetic evaluation, the three evaluation scores of a paragraph in a news story must be first fuzzified as fuzzy degrees based on the fuzzy membership functions automatically determined in K-means clustering. Define the centers of three linguistic terms determined by the K-means clustering algorithm as c_1, c_2, and c_3 for the "length score" indicator. Fig. 4 illustrates an example of fuzzy membership function automatically determined using the K-means clustering algorithm for the indicator "length score".

The detailed steps of determining the representativeness of a paragraph in a Google news story based on fuzzy synthetic evaluation are presented as follows:

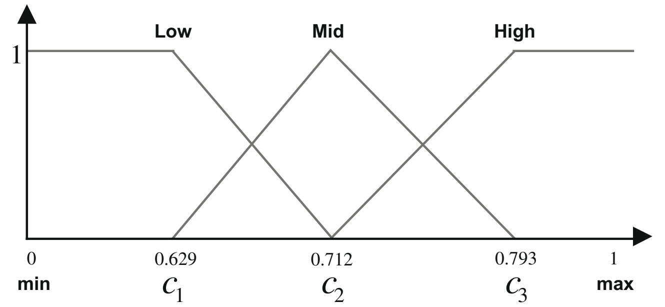


Fig. 4. An example of the fuzzy membership functions determined by the centers of three linguistic terms for the indicator "length score".

Step 1: Construct the evaluation set of the indicators for news summarization
The evaluation set of representativeness indicators of a paragraph are constructed, and represented as $\underset{\sim}{U} = [\text{LS} \quad \text{SS} \quad \text{IS}]$.

Step 2: Determine the corresponding weights of the considered indicators in the evaluation set.
Weights are assigned to each considered evaluation indicator depending on its importance in news summarization identification. The weight design is strongly affects the results of the fuzzy synthetic evaluation method. The weight combination of the considered evaluation indicators is based on decisions by language experts. The weight matrix can be represented as $\underset{\sim}{W} = [\text{LS}_{\text{weight}} \quad \text{SS}_{\text{weight}} \quad \text{IS}_{\text{weight}}]$. For example, consider the case where the weight of each indicator is 0.3, 0.5, and 0.2, respectively. This example emphasizes similarity as the most important indicator, followed by length and information.

Step 3: Assign the evaluation fuzzy linguistic terms for each considered indicator.
The fuzzy linguistic terms for each considered indicator in the evaluation set are then computed. To simplify the fuzzy inference procedure, the evaluated linguistic fuzzy terms for each considered indicator are determined as $\underset{\sim}{V}_i = [\text{Low} \quad \text{Moderate} \quad \text{High}]$ in this study, where $\underset{\sim}{V}_i$ denotes the fuzzy term set of the ith considered indicator, and i = 1, 2, and 3.

Step 4: Derive fuzzy relation matrix based on the fuzzy membership functions determined by K-means clustering for three considered evaluation indicators of news summarization identification.
The crisp score values assessed by three evaluation scores are fuzzified into fuzzy degrees in order to obtain the fuzzy relation matrix based on the corresponding fuzzy membership functions as determined by the K-means clustering scheme. For example, consider the following fuzzy relation matrix for identifying the representativeness of a paragraph:

$$\underset{\sim}{R} = \underset{\sim}{U} \times \underset{\sim}{V} = \begin{bmatrix} 0.2 & 0.5 & 0.2 \\ 0.7 & 0.2 & 0.1 \\ 0.1 & 0.7 & 0.2 \end{bmatrix} \tag{7}$$

Step 5: Obtain the result of fuzzy synthetic evaluation based on fuzzy matrix composition operation.
The assessing result in the fuzzy synthetic evaluation method can be reasoned using various methods, such as fuzzy operator, weighted average, distance and additive operation (Chang et al., 2001). This study adopted the fuzzy operator method to reason the assessing result. The fuzzy operator method performs the fuzzy synthetic evaluation using the maximum–minimum operator, and the final assessing result is dominated by the most influential factor. The reasoning process for this example is given as follows:

$$\underset{\sim}{Y} = \underset{\sim}{W} \circ \underset{\sim}{R} = [0.3 \quad 0.5 \quad 0.2] \circ \begin{bmatrix} 0.2 & 0.5 & 0.2 \\ 0.7 & 0.2 & 0.1 \\ 0.1 & 0.7 & 0.2 \end{bmatrix} = [0.5 \quad 0.3 \quad 0.2] \tag{8}$$

Step 6: Defuzzifying for obtaining the important degree of a paragraph
Defuzzification means calculating the crisp value of a fuzzy number. The crisp value approximately represents the deterministic properties of the fuzzy reasoning process according to the assessment matrix, and helps convert the uncertainty into an applicable action when solving real-world problems. Suppose that different defuzzification scales are assigned to three considered indicators in order to infer the importance degree of a paragraph. For example, assign the scale $\lambda_1 = 1$ to the length score V_{i1}; $\lambda_2 = 3$ to the similarity score V_{i2}, and $\lambda_3 = 5$ to the information score V_{i3}. In this case, the inferred result is approximately 0.27 after performing defuzzification. The detailed mathematical formula for defuzzification is given below.

$$q = \frac{\sum_{i=1}^{3} y_i \times \lambda_i}{\sum_{i=1}^{3} \lambda_i} = \frac{0.5 \times 1 + 0.3 \times 3 + 0.2 \times 5}{1 + 3 + 5} \cong 0.27 \tag{9}$$

4. Proposed user location detection scheme based on GPS signals

This section describes the proposed user location detection scheme to identify the current position in Taiwan of a user based on the longitude and latitude coordinates sensed by GPS. The following subsections detail the proposed scheme.

4.1. Problem description

An intelligent location-based news service system must precisely identify a user's location from longitude and latitude coordinates. In this issue, the GIS group of Academia Sinica in Taiwan has developed a GIS research supported tool that can locate a point in any town in Taiwan (http://gis.ascc.net/ISTIS/program/PntInTw/PntInTw.zip) from recorded vector-based polygon data. The tool first identifies the polygon in which the user is located based on longitude and latitude coordinates. The user location is then identified, because each polygon has a corresponding location label. However, since many polygon data need to be recorded to locate a user, running the appropriate algorithm on a PDA causes a high computational load. To solve this problem, this study presents a scheme based on machine-learning to reason the location of a user from a few measured sampling points in a location area. This study adopted the longitude and latitude coordinates to locate sampling point. The centers of city districts were utilized as sampling points. For example, Taipei city has 12 sampling points for identifying the user location, because it has 12 districts. Fig. 5 shows the sampling of representative points to locate users in Taipei city by the proposed location identification scheme based on machine-learning.

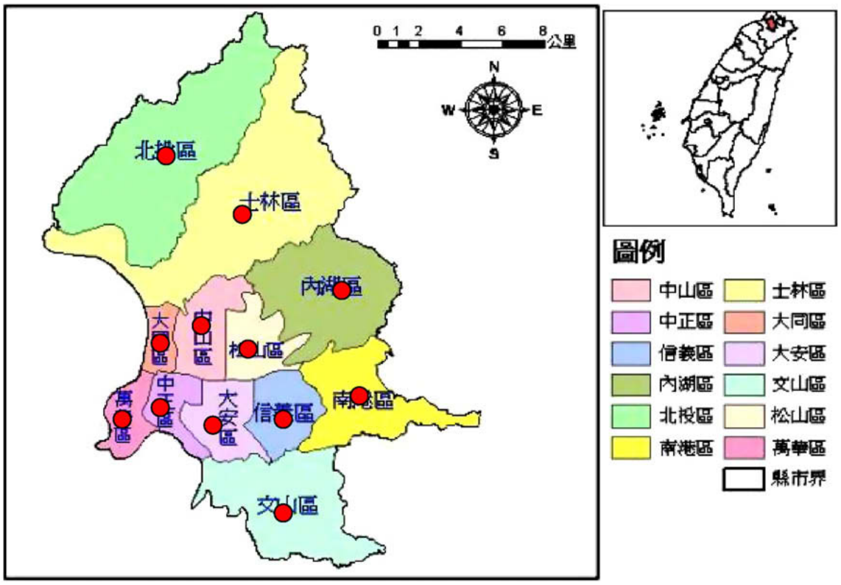


Fig. 5. An example for showing how to sample representative points for identifying user location by the proposed location identification scheme (the red marks represent the sampling points).

4.2. The proposed average k-NN classifier for identifying user location

Detecting the location of a user is important in the proposed intelligent location-based news service system. After analyzing the advantages and disadvantages of several positioning techniques, and considering the limitation of the real environment into account, this study develops a positioning method based on the average *k*-NN classifier, which is a variant of *k*-nearest neighbor classifier, to develop a location-based news service based on the GPS signals. In pattern recognition, the *k*-nearest neighbor classifier (*k*-NN) classifies objects according to the closest training examples in the feature space. *k*-NN is an instance-based learning system, and is also-called lazy learning, since the function is only approximated locally, and all computation is deferred until classification. A major limitation of using this algorithm to classify a test pattern is that the classes with more frequent training samples tend to dominate the prediction of the test pattern, because they tend to appear in the *k*-nearest neighbors when the neighbors are calculated due to their large number. This problem can be solved by considering the distance of each *k*-nearest neighbors with the test pattern that is to be classified, and predicting the class of the test pattern from these distances (Wettschereck, Aha, & Mohri, 1997).

Therefore, this study proposes the average *k*-NN classifier, which considers the distance of each of *k*-nearest neighbors with the test pattern, to reason user location based on the gathered data pairs comprising the longitude coordinate, latitude coordinate and corresponding location label. The data pairs were obtained from the geography information database provided by Ministry of The Interior of Taiwan Government. In contrast to the original *k*-NN classifier, the proposed average *k*-NN classifier addresses the average distances between the test pattern with the *k*-nearest neighbors belonging to different classes, rather than classifying by a majority vote of the *k*-nearest neighbors with the test pattern. Hence, the test pattern is recognized as the class with the shortest average distance while applying the proposed average *k*-NN classifier. The proposed average *k*-NN classifier can be formulated as follows:

$$\hat{f}(x_q) \leftarrow \arg\min_{v \in V} \frac{\sum_{i=1}^{k} dw_i \delta(v, f(x_i))}{\sum_{i=1}^{k} \delta(v, f(x_i))} \tag{10}$$

$$\delta(a,b) = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

$$dw_i = \sqrt{x_i^2 - x_q^2} \tag{12}$$

where $\hat{f}(x_q)$ denotes the class predicted by pattern x_q; $f(x_i)$ represents the class of the *i*th neighbor pattern of the pattern x_q; v is the set of all classes; x_i is the *i*th neighbor pattern of the pattern x_q, and x_q is the pattern to be classified.

Additionally, the method of selecting *k* is an essential issue in the proposed average *k*-NN classifier, because it affects the classification accuracy rate. The optimal value of *k* depends on the data distribution. Generally, larger *k* decreases the effect of noise on the classification, but makes boundaries between classes less distinct. A good *k* can be selected using various heuristic or optimal techniques, such as cross-validation and genetic algorithm (Li, Lu, & Yu, 2004). An example is given below to explain how to determine the class label for identifying a user located at Taiwan city according to the proposed average *k*-NN classifier. Fig. 6 illustrates the example for identifying user location by the proposed average *k*-NN classifier. In this example, suppose the number of neighbor patterns is set to 9 (i.e. *k* = 9). Among the nine nearest neighbors with the test pattern, two training samples belong to class A, and seven training samples belong to the class B. The distances between the

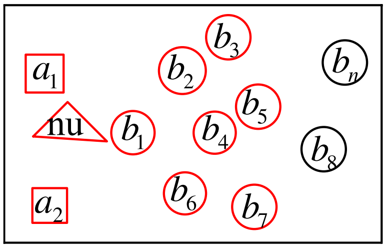


Fig. 6. An example for identifying user location by the proposed average *k*-NN classifier.

test pattern and the training samples in class A are 1 and 2, and the distances between the test pattern and the training samples in class B are 0.5, 3, 4, 3.5, 4, 3 and 4. Therefore, the average distance between the test pattern and the two neighbors in class A can be computed as

$$A\_\text{averager}\_\text{dist} = \frac{1+2}{2} = 1.5 \tag{13}$$

Similarly, the average distance between the test pattern with the seven neighbors belonging to class B can be computed as

$$B\_\text{averager}\_\text{dist} = \frac{0.5+3+4+3.5+4+3+4}{7} = 3.1 \tag{14}$$

Since the proposed average *k*-NN classifier classifies the test pattern according to the shortest average distance, the test pattern is classified in class A in this example. Later experiments indicate that the proposed average *k*-NN classifier is clearly superior to the four well-known machine-learning schemes in the solving location identification problem, especially for identifying boundary locations between two neighbor cities.

5. Experiments

This section first presents the results of quality evaluation of the proposed multi-document news summarization scheme based on manual judgment. To demonstrate the performance of the proposed user location identification scheme, the positioning accuracy rates of the proposed average *k*-NN classifier were assessed by various well-known machine-learning schemes based on some randomly selected test samples from UrMap (2009). Finally, the implementation of the location-based news service with automatic news summarization on PDA mobile devices is described.

5.1. Evaluation method of generating news summarization

To assess the quality of the generated news summaries, 71 pieces of news stories categorized in the class of “Taiwan social news” were randomly retrieved from the Google news site as testing samples to evaluate the performance of the proposed multi-document news summarization scheme. Each sampled news story groups together related news articles on the same topic but reported by different media. The generated summary is hard to evaluate automatically, because the “ideal summary” depends on the user, purpose and genre. Previous studies have often used summary evaluation methods, including manual judgment, similarity evaluation, task-based evaluation and questionnaire evaluation, as indicated in our literature survey (Mani, 2001). Among these summary evaluation methods, manual judgment evaluation involves inviting people to judge the generated summary based on several designed rating scales. Although this method is labor-intensive and time-consuming, it is the most direct and accurate if the test and guidelines are properly configured (Nanba & Okumura, 2006). Therefore, this study adopted manual judgment evaluation with three evaluation scales to assess the quality of a generated news summary. The three evaluation scales that were employed to measure the quality of news summaries are explained as follows:

(1) Good: the generated summary is representative, sufficient information, and is of appropriate length.
(2) Ok: the generated summary is representative, but the information is insufficient or the length is inappropriate.
(3) Bad: the generated summary is not representative.

5.2. Quality evaluation of generating news summarization

The impacts of different weight matrices influence the news summarization quality generated by the employed fuzzy synthetic evaluation method. To simplify the evaluation results, the rating results of “good” and “Ok” were merged into “satisfactory”, and the rating result of “bad” was classified as “unsatisfactory”. The summarization quality of news stories randomly chosen from Google news class was analyzed. Table 1 shows the quality evaluation results of news summarization of 71 randomly selected Google news stories in the class “Taiwan social news” based on five language experts under different 11 weight matrices. In the experiment, the generated news summary had the highest “satisfactory” rating then the weights of the length, similarity, and information scores were respectively set as 0.2, 0.4 and 04. Observation results indicate that the weight combinations of the similarity and information scores are more important indicators of news summarization quality than the length score. Experimental results show that assigning an excessive weight to the length score leads to poor news summarization quality. The satisfactory rating decreased with increasing weight of the length score. This is because an over-long news summary leads to inconvenient reading. Therefore, this study confirms that assigning approximately equal weights for the similarity and information scores, while assigning a relatively low weight for the length score, leads to high-quality news summaries.

Finally, the position distribution of the generated news summarization paragraphs was analyzed according to these experimental results. Table 2 lists the appearance positions of summary paragraphs, along with their quality ratings. The proportions of the generated news summaries appearing in the first, second and last paragraphs were 44.53%, 26.07% and 11.15%, respectively. Only 18.27% of the news summaries in the remaining positions. The results confirm that most news writing structures belong to the

Table 1
The quality evaluation results of news summarization of 71 randomly selected Google news categorized in the class of “Taiwan social news” based on five language experts under different 11 weight matrices.

Experiment	Quality					
	LS weight	SS weight	IS weight	Good	Ok	Bad
1	0.2	0.7	0.1	49(69%)	19(26.7%)	3(4%)
2	0.2	0.5	0.3	58(81.7%)	10(14.1%)	3(4%)
3	0.2	0.4	0.4	53(74.6%)	17(23.9%)	1(1.4%)
4	0.2	0.3	0.5	54(76.1%)	15(21%)	2(2.8%)
5	0.2	0.1	0.7	43(60.6%)	26(36.6%)	2(2.8%)
6	0.4	0.1	0.5	44(62%)	21(30%)	6(8.5%)
7	0.4	0.2	0.4	50(70%)	16(22.5%)	5(7%)
8	0.4	0.3	0.3	51(71.8%)	16(22.5%)	4(5.6%)
9	0.4	0.4	0.2	54(76.1%)	12(17%)	5(7%)
10	0.4	0.5	0.1	54(76.1%)	13(18%)	4(5.6%)
11	0.6	0.2	0.2	51(71.8%)	11(15.5%)	8(11.3%)
	Average			71.8%	20.35%	5.2%

Table 2
The statistical information of the summarization paragraph positions for the news summarization with satisfied rating levels.

Experiment	Position			
	The first paragraph	The second paragraph	The last paragraph	Others
1	39(57.35%)	13(19.12%)	5(7.35%)	11(16.18%)
2	32(47.06%)	16(23.53%)	6(8.82%)	14(20.59%)
3	30(42.86%)	19(27.14%)	10(14.29%)	11(15.71%)
4	38(55.07%)	17(24.64%)	10(14.45%)	4(5.76%)
5	37(53.62%)	18(26.09%)	4(5.8%)	10(14.49%)
6	27(41.53%)	21(32.31%)	4(6.15%)	13(20%)
7	20(33.9%)	15(25.42%)	9(15.25%)	15(25.42%)
8	24(36.36%)	21(31.82%)	7(10.61%)	14(21.21%)
9	28(42.42%)	19(28.79%)	7(10.61%)	12(18.18%)
10	33(49.25%)	14(20.9%)	9(13.43%)	11(16.42%)
11	19(30.16%)	17(26.98%)	10(15.87%)	17(26.98%)
Average	44.53%	26.07%	11.15%	18.27%

so-called "inverted pyramid", in which the significance of information declines as the article progresses (Marcus et al., 2009). In other words, news writing attempts to answer all the basic questions about any particular event in the first two paragraphs.

5.3. Accuracy evaluation of the proposed user location detection scheme

The performance of identifying user's location for the proposed average *k*-NN classifier was compared with various machine-learning schemes, namely BP neural networks, naïve Bayesian classifier, support vector machine and original *k*-NN classifier. In the experiment, the performance of all test schemes in identifying user locations was measured by 40 randomly selected user positions from UrMap (2009). Half of the position samples were located at the neighborhood area of 20 city centers, and the remaining half were located at the boundary area between cities in Taiwan. The accuracy of each of these five machine-learning schemes was then reported.

5.3.1. BP neural networks

First, BP neural networks (Rumelhart, Hiton, & Williams, 1996) with different learning structures and parameter settings were used to evaluate the performance of identifying user locations. Experimental results indicate that the BP neural networks could not solve this problem well under different learning structures with parameter selection. Fig. 7 shows the corresponding mean square errors of the BP neural networks with various numbers of hidden neurons when setting both the learning rate and momentum to 0.5. In this experiment, the mean square error did not converge to a satisfactory learning error, even the number of hidden neurons was increased from 1 to 16. Fig. 8 shows the accuracy rates of the BP neural networks with various numbers of hidden neurons for identifying neighborhood locations of city center and boundary locations between cities. The results in this figure reveal that user location identified by the BP neural networks could not achieve a satisfactory accuracy rate, regardless of the identifying neighbors of city center or city boundary.

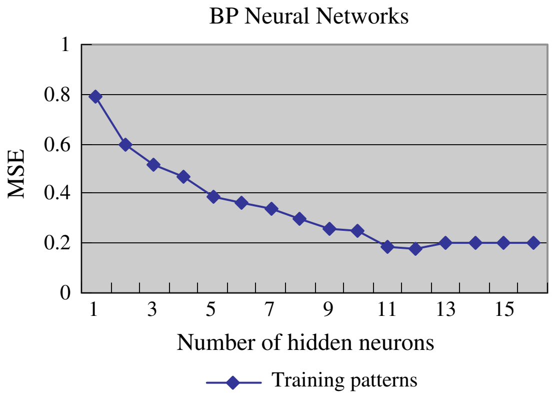


Fig. 7. The corresponding mean square errors of the BP neural networks with various numbers of hidden neurons under fixed input and output nodes.

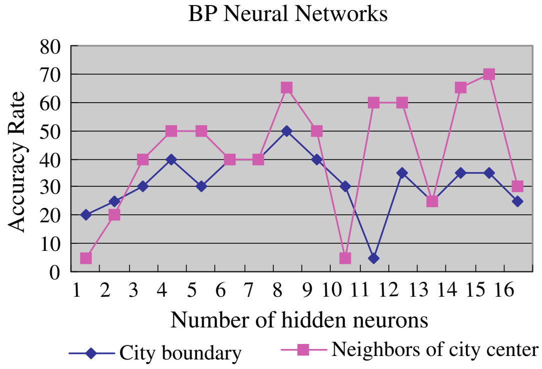


Fig. 8. Accuracy rate of user location-awareness identified by BP neural networks with various numbers of hidden neurons for neighborhood locations of city center and boundary locations between cities.

5.3.2. Naïve Bayesian classifier

The naïve Bayesian classifier was then run to solve the same user location identification problem. A naïve Bayesian classifier is a simple probabilistic classifier that applies Bayes' theorem with naïve independence assumptions. The naïve Bayesian classifier is one of the most efficient and effective inductive learning algorithms for machine-learning and data mining, and performs very well in classification (Dunham, 2002). However, it is based on

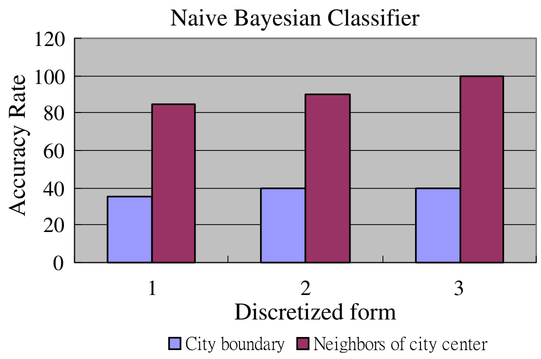


Fig. 9. Accuracy rates of the naïve Bayesian classifier with different partition forms.

the conditional independence assumption that almost never holds, thus possibly leading to poor performance. Additionally, non-discrete features need to be discretized first when adopting naïve Bayesian classifier for pattern classification. Three discretized forms were performed based on *ad hoc* selection of bins to discretize input features into a series of intervals. Fig. 9 shows accuracy rates of the naïve Bayesian classifier with different discretized forms for identifying user locations. The experimental results indicate that although the third discretized forms achieve an accuracy rate of up to 100% for identifying neighbors of a city center, it has a very poor accuracy rate for identifying city boundaries.

5.3.3. Support vector machine

Lin's support vector machine (LIBSVM) (2009) was also run to solve the user location identification problem. The LIBSVM can automatically compute the SVM parameters including the used kernel (default as the radial basis function) and kernel parameters using grid parameter searching (LIBSVM, 2009). In the experiment, although Lin's LIBSVM achieved an accuracy rate of up to 95% when identifying neighbors of a city center, it 60% accuracy for the user positions located at a city boundary. Obviously, Lin's LIBSVM cannot effectively solve non-linear pattern classification problems like this case.

5.3.4. Original k-NN algorithm with the proposed average k-NN classifier

The most important parameter in a text categorization system based on the *k*-nearest neighbor algorithm is *k* (Li et al., 2004). Fig. 10 shows the comparison of accuracy rates for identifying city boundary utilizing the average *k*-NN classifier with the original *k*-NN classifier under different selected *k* values. Regardless of the value of *k*, the accuracy rate of the proposed average *k*-NN classifier was greater than or equal to 70%, while that of the original *k*-NN classifier was less than or equal to 70%. The proposed average *k*-NN classifier for identifying user location located at city boundary obviously has a superior accuracy rate than the BP neural networks, naïve Bayesian classifier, support vector machine, and original *k*-NN classifier. Fig. 11 shows the overall accuracy rates for identifying neighbors of city center and city boundary using the average *k*-NN classifier with the original *k*-NN classifier under different *k* values. The overall accuracy rate of the proposed average *k*-NN classifier, unlike the original *k*-NN classifier, was higher than 80% regardless of the *k* values.

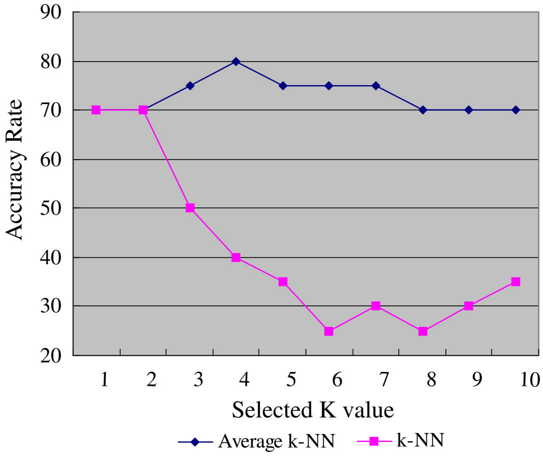


Fig. 10. Comparison of accuracy rate of city boundary for the average *k*-NN with the original *k*-NN classifier.

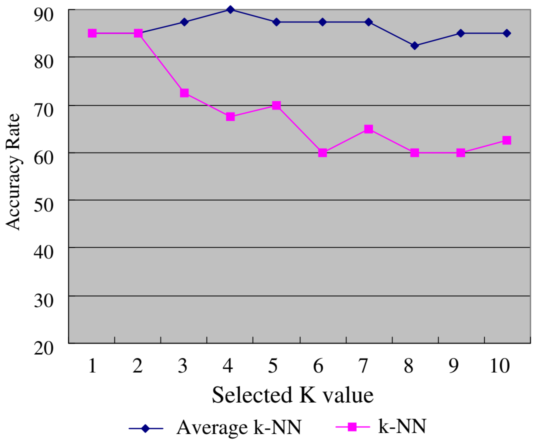


Fig. 11. Comparison of entire accuracy rates including neighbors of city center and city boundary for the average *k*-NN with the original *k*-NN classifier under different *k* values.

5.4. Accuracy evaluation results of user location identification in the tested machine-learning schemes

Evaluation results of these tests indicate that the BP neural networks, naïve Bayesian classifier, and SVM did not achieve a satisfactory accuracy rate for identifying user location problem. The first likely reason is the skewed distribution of sampling data. Large cities contain much more sampling data than small cities, causing the machine-learning schemes to fit for the sampling training data in large cities, thereby poorly fitting the sampling training data in small cities. The second likely reason is that fitting the sampling data located at an irregular city boundary is very challenging to any machine-learning technologies. Therefore, the naïve Bayesian classifier with linear pattern recognition capability cannot solve this non-linear problem well. Similarly, the support vector machines (SVMs) also belong to a family of generalized linear classifiers. A special property of SVMs is that they simulta-

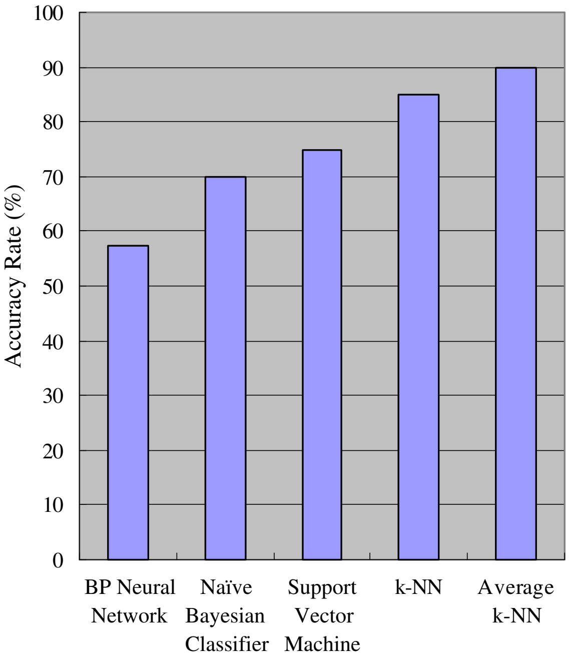
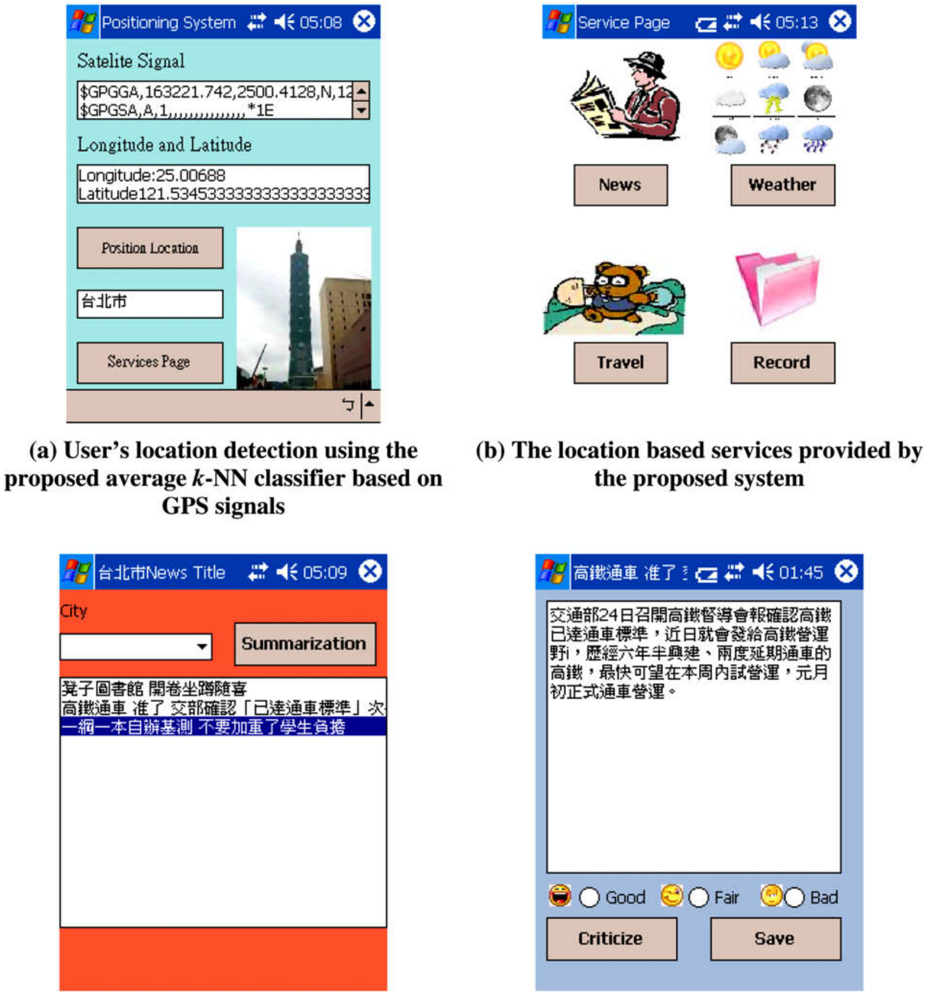


Fig. 12. Comparison of the accuracy rate of user location detection for the proposed average *k*-NN classifier with various classifiers.



$GPGGA,163221.742,2500.4128,N,12
$GPGSA,A,1,,,,,,,,,,,,,,,,*1E

Longitude:25.00688
Latitude121.534533333333333333333

台北市

(a) User's location detection using the proposed average *k*-NN classifier based on GPS signals

(b) The location based services provided by the proposed system

凳子圖書館 開卷坐蹲隨喜
高鐵通車 准了 交部確認「已達通車標準」次
一綱一本自辦基測 不要加重了學生負擔

(c) The dispatching location-based news based on user's location

交通部24日召開高鐵督導會報確認高鐵已達通車標準，近日就會發給高鐵營運野i，歷經六年半興建、兩度延期通車的高鐵，最快可望在本周內試營運，元月初正式通車營運。

○ Good ○ Fair ○ Bad

(d) The news summarization result for a selected news

Fig. 13. The implemented location-based news service with automatic news summarization on PDA.

neously minimize the empirical classification error and maximize the geometric margin, and thus are also known as maximum margin classifiers (Burges, 1998). In these experiments, SVM also did not achieve a satisfactory accuracy rate in solving the user's location identification problem. An essential issue in SVM is that the pattern classification accuracy rate is obviously affected by the parameters of the selected kernel function. Meanwhile, appropriately determining these parameters is also a critical research issue. Although BP neural networks are non-linear classifiers, these experiments show they cannot predict the solved user location identification problem well due to the likely skewed distribution of training samples. The original *k*-NN algorithm solved the problem better than the BP neural network, naïve Bayesian classifier and SVM, according to experimental results, but it easily misjudges the user's location in the case of large cities neighboring small cities, because the more frequent training samples tend to dominate the prediction of the test pattern. Therefore, the average *k*-NN classifier is presented to overcome this problem, since it considers the distance of each *k*-nearest neighbors to the test pattern. Fig. 12 compares the accuracy rates of user location detection for various classifiers with the proposed average *k*-NN classifier. Experimental results confirm that the proposed average *k*-NN classifier has the best performance in terms of identifying user's location among the five tested machine-learning schemes.

5.5. Implemented system for location-based news service on PDA device

The proposed intelligent location-based news service system with automatic news summary was implemented using Microsoft Visual Basic .NET 2003. The software tool can support over two hundred web mobile devices, including mobile phone and PDA, to promote the development speed of applications. It is suitable for developing intelligent devices like Pocket PC. Fig. 13(a) shows the user's location detection using the proposed average *k*-NN classifier based on GPS signals. In this example, the identified user position is Taipei city based on satellite signals from GPS. Fig. 13(b) shows the location-based services provided by the proposed system other than providing location-based news service. However, this study focuses only on the news service. Fig. 13(c) illustrates the dispatching location-based news based on user location, and Fig. 13(d) shows the news summarization result for a piece of news selected by a user.

6. Conclusion

This study presents an intelligent location-based news service system with automatic news summarization based on the proposed multi-document news summarization scheme and average *k*-NN classifier. The proposed scheme can accurately extract one most representative paragraph as a summary of a news story that contains news articles reported by different media from the Google news site. The generated news summary is then immediately dispatched to individual users with PDA mobile devices based on the user location identified by the proposed average *k*-NN classifier based on GPS signals. In real-world applications, the proposed system conveniently enables users to obtain everyday local news associated with their current locations. The proposed system accelerates the valuation of news events, since the news events often have time validity. Evaluation results reveal that the proposed system — based on setting appropriate weight matrix for the fuzzy synthetic evaluation method — obtains a satisfactory news summarization quality while identifying a most representative news summarization paragraph. Additionally, the proposed average *k*-NN classifier for identifying the location of a user in Taiwan area is better than four well-known machine-learning schemes tested in this study.

References

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*, 121–167.

Chang, N.-B., Chen, H. W., & Ning, S. K. (2001). Identification of river water quality using the fuzzy synthetic evaluation approach. *Journal of Environmental Management, 63*(3), 293–305.

Chen, C.-M., & Liu, C.-Y. (2009). Personalized e-news monitoring agent system for tracking user-interested news events. *Applied Intelligence, 30*(2), 121–141.

Chen, C. -M., Li, Y. -L., & Chen, M. -C. (2007). Personalized context-aware ubiquitous learning system for supporting effectively English vocabulary learning. In *IEEE international conference on advanced learning technologies* (pp. 628–630).

Chen, H.-H., Kuo, J.-J., Huang, S.-J., Lin, C.-J., & Wung, H.-C. (2003). A summarization system for Chinese news from multiple sources. *Journal of the American Society for Information Science and Technology, 54*(13), 1224–1236.

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*.

Choi, D.-Y. (2007). Personalized local internet in the location-based mobile web search. *Decision Support Systems, 43*, 31–45.

Dunham, M. H. (2002). *Data mining: Introductory and advanced topics*. Prentice Hall Inc..

ECScanner (An Extension Chinese Lexicon Scanner). (2009). Available from: http://demo.lias.nccu.edu.tw:58080/rank/www/ecscanner/.

Fung, G. P. C., Yu, J. X., & 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).

Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. (1999). Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of the 22nd international ACM conference on research and development in information retrieval (SIGIR-99)*, Berkeley, CA, pp. 121–128.

Google news. (2009). Available from: http://www.google.com/press/descriptions.html#news.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient *k*-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(7), 881–892.

Kuo, Y.-F., & Chen, P.-C. (2006). Selection of mobile value-added services for system operators using fuzzy synthetic evaluation. *Expert Systems with Applications, 30*(4), 612–620.

Li, B. L., Lu, Q., & Yu, S. W. (2004). An adaptive *k*-nearest neighbor text categorization strategy. *ACM Transactions on Asian Language Information Processing, 3*(4), 215–226.

LIBSVM. (2009). Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Mani, I. (2001). Summarization evaluation: An overview. In *Proceedings of the NAACL 2001 workshop on automatic summarization*.

Marcus, E. et al. (2009). The evolution of the summary news lead. Available from: http://www.scripps.ohiou.edu/mediahistory/mhmjour1-1.htm.

Nanba, H., & Okumura, M. (2006). An automatic method for summary evaluation using multiple evaluation results by a manual method. In *Proceedings of the COLING/ACL on main conference poster sessions* (pp. 603–610).

Otterbacher, J., Radev, D., & Kareem, O. (2008). Hierarchical summarization for delivering information to mobile devices. *Information Processing and Management, 44*(2), 931–947.

Radev, D. R., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization. *Computational Linguistics, 28*(4), 399–407.

Rashid, O., Mullins, I., Coulton, P., & Edwards, R. (2006). Extending cyberspace: Location based games using cellular phones. *Computers in Entertainment, 4*(1), 1–18.

Rodriguez, M. D., Favela, J., Martinez, E. A., & Munoz, M. A. (2004). Location-aware access to hospital information and services. *IEEE Transactions on Information Technology in Biomedicine, 8*(4), 448–455.

Rumelhart, D. E., Hiton, G. E., & Williams, R. J. (1996). Learning internal representation by error propagation. *Parallel Distributed Processing, 1*, 318–362.

Salton, G. (1989). *Automatic text processing – The transformation, analysis, and retrieval of information by computer*. Addison-Wesley.

Salton, G., & McGill, J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill Book Company.

UrMap. (2009). Available from: http://www.urmap.com/.

Virrantaus, K., Markkula, J., Garmash, A., & Terziyan, Y. V. (2001). Developing GIS-supported location-based services. In *Proceedings of WGIS'2001 – First international workshop on web geographical information systems*, Kyoto, Japan, pp. 423–432.

Wettschereck, D., Aha, D. W., & Mohri, T. (1997). A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review, 11*(1–5), 273–314.

Yang, C. C., & Wang, F. L. (2007). An information delivery system with automatic summarization for mobile commerce. *Decision Support Systems, 43*(1), 46–61.

Yoshino, T., Muta, T., & Munemori, J. (2002). NAMBA: Location-aware collaboration system for shopping and meeting. *IEEE Transactions on Consumer Electronics, 48*(3), 470–477.