J Med Syst (2013) 37:9946
DOI 10.1007/s10916-013-9946-3

ORIGINAL PAPER

Web Information Retrieval for Health Professionals

S. L. Ting · Eric W. K. See-To · Y. K. Tse

Received: 5 February 2013 / Accepted: 25 March 2013

Abstract This paper presents a Web Information Retrieval System (WebIRS), which is designed to assist the healthcare professionals to obtain up-to-date medical knowledge and information via the World Wide Web (WWW). The system leverages the document classification and text summarization techniques to deliver the highly correlated medical information to the physicians. The system architecture of the proposed WebIRS is first discussed, and then a case study on an application of the proposed system in a Hong Kong medical organization is presented to illustrate the adoption process and a questionnaire is administrated to collect feedback on the operation and performance of WebIRS in comparison with conventional information retrieval in the WWW. A prototype system has been constructed and implemented on a trial basis in a medical organization. It has proven to be of benefit to healthcare professionals through its automatic functions in classification and summarizing the medical information that the physicians needed and interested. The results of the case study show that with the use of the proposed WebIRS, significant reduction of searching time and effort, with retrieval of highly relevant materials can be attained.

Keyword Web information retrieval · Text mining · Web mining

Introduction

Knowledge is synthesized and organized information that help explain events and situations. Knowledge also helps to evaluate and make connections between current and future happening, so as to facilitate decision makings [1]. In particular within the medical field, knowledge is important to medical professionals to make clinical judgments and decisions, so as to help improve the quality of care and services provided to patients. For example, health professionals have to rely on their knowledge, and then connect them with the patients' symptoms. Thereby, they can come up with a conclusion with the patients' disease and prescribe suitable diagnosis and medicines.

The World Wide Web (WWW) is one of the useful and common sources for medical professionals to acquire and obtain up-to-date knowledge [2]. According to the result from the annual interview conducted by the American Medical Association (AMA) in 2007, the number of American physicians who are Internet users have significantly increase in the previous 10 years, from 35 % of surveyed physicians in 1997 to 94 % surveyed in 2007, and it is also estimated that the result will continuously demonstrate an upward trend.

In the web, health professionals may browse through news articles, medical journals and other codified materials. The need of updated medical information and knowledge are especially vital to health professionals [3–7]. Take the case of Swine Flu as an example, since Swine Flu is a newly discovered disease, medical professionals have little knowledge on how to cure it and how should the medical therapy be. As a result, they have to depend on the web, which aid efficient and convenient exchange and sharing of knowledge without time and geographical boundary, to capture the latest information.

However, there are millions of information and sources available in the web. They might be duplicated or even irrelevant to what the medical professionals are looking for when they are searching for it through a search engine. Thus, medical professionals are facing the problem of information overload [8]. In order to sort out the necessary piece of information, physicians have to go through the search results one after another. A long processing time and great

S. L. Ting (✉) · E. W. K. See-To
Department of Industrial and Systems Engineering,
The Hong Kong Polytechnic University, Hung Hom,
Kowloon, Hong Kong
e-mail: jacky.ting@connect.polyu.hk

Y. K. Tse
The York Management School, University of York,
Freboys Lane, Heslington York YO10 5GD, UK

Published online: 10 April 2013

effort are needed to find out the relevant web information which suits their needs. Furthermore, another challenge that physicians encountered in retrieving web information is information timeliness. As medical information is time-sensitive, clinical facts that once were true can easily be replaced by updated information and development [9]. Therefore, physicians are keen to obtain and rely on the latest knowledge to provide quality care and diagnosis to patients.

With the aim to provide healthcare professionals with the latest sources of relevant medical knowledge, this paper intended to propose a Web Information Retrieval System (WebIRS) to help improve the situation mentioned above. The proposed system is expected to make use of text and web mining techniques, which include document classification and text summarization, to aid health professionals' information retrieval processes in the WWW. In order to investigate the possibility and usefulness of the proposed approach within the health care domain, a Hong Kong medical organization is used as a reference case company for the consultation in system development, as well as the methodology evaluation.

Literature review

Information retrieval and query

Information retrieval means finding material, mostly documents, of an unstructured nature (usually text) that satisfies information need from large collections [10]. It is the science of searching for documents, for information, for metadata and searching the WWW [11]. Due to the increasing influence of technology and the WWW, millions of people are engaging in information retrieval activities everyday when they use a search engine or even to search their email. Information retrieval has become the dominant form of information access nowadays.

According to Manning et al. [10], information retrieval system aims to provide documents from within the collection that are relevant to an arbitrary user information need, communicated to the system by means of a one-off, user-initiated query. An information query is what the user conveys to the computer in an attempt to communicate the information need, and is differentiated from an information need, which is about the topic which the user desires to know more.

According to numerous studies [10, 12], the common web search queries can be grouped into three groups, which are informational, navigational and transactional. Informational queries seek general information in a broad topic. There is typically no single web page that contains all the information required. As a result, users with informational queries usually try to assimilate information from multiple web pages. The second group is navigational queries in which this type of query seeks the websites of a single entity that the user has in mind. This type of users would expect the very first search result should be the page that he/she is looking for. The user is not interested in other materials that contain the information regarding that single entity. The last group is the transactional query which is a prelude to the user performing a transaction on the Web, such as purchasing a product, downloading a file, or making a reservation. In such case, the user is expecting the search engine to return results listing services that provide form interfaces for such transaction.

In developing a web information retrieval system, it is important to understand the users' query need [13]. It is obvious that some users' queries might fall into more than one of these groups, whereas some might fall outside them. However, it is argued that the queries inputted by physicians would be mainly the informational one, and with little chance that the queries would be transactional query.

Importance of healthcare information from web

WWW is one of the useful sources for the medical domain and the health professionals to obtain the latest medical knowledge and information. The significance of the WWW and the Internet to physicians' professional development is growing quickly. According to Bell and Sethi [2], many physicians are now utilizing the WWW to access medical journal articles in place of traditional print journals. Physicians claimed that the WWW offers them with a channel for quick and 24-h access to information; also, its ease of searching is another reason for switching to online information.

In accordance to a survey conducted by the AMA [8], Internet's importance to physicians nowadays is mainly in two areas; first, for information seeking, and second, for professional development through online continue medical education. Interviewed physicians claimed that, they major purpose for utilizing internet to search for information is to solve a particular patient's problem and to seek for providing better care to patients.

It is crucial for physicians to become aware of the latest medical finding. It is found that by delivering the newly discovered techniques, therapy formula and medication to patients; it can significantly improve the accuracy of diagnosis and the quality of care. For example, new information about HIV/AIDS is emerging more rapidly than any other disease in history and new pharmaceuticals for this are being created everyday [14]. However, clinical information is time-sensitive. Many clinical facts once were true can easily being replaced by updated and new information [9]. With the help of the WWW, the diffusion of new medical knowledge to

healthcare professionals could be fastened. When clinicians' awareness of new diagnostic and treatment knowledge continue to increase, their care and service quality will ultimately be improved. As a result, many physicians now use internet to keep themselves up-to-date.

Current approaches and challenges in web healthcare information retrieval

Nowadays, nearly all physicians have access to the Internet, and know how to use it for accessing medical information [8]. The most common way for physicians to obtain their required information is to type in a keyword or a phrase as the search query into a search engine, and then wait for a list of research results in return. In 2009, Morehouse School of Medicine has conducted a survey to collect physicians' preferred sources of information [15]. Their study reviewed that the top three physicians' preferred medical information sources are professional journals, medical websites, and professional association updates. It is, as well, found that physicians prefer to obtain healthcare information from professional and creditable sources. As a result, they would mostly browse the journals, websites, and corresponding links which are published in authoritative medical web pages.

As the quality of service delivered by physicians are highly dependent on the availability of accurate information, it is important for them to retrieve information which is only relevant to current patient treatment [16]. However, physicians are usually with heavy workload schedules, it is impossible for them to validate every piece of information retrieved from the web. Moreover, it is particularly at risk for physicians to have irrelevant, incomplete or outdated knowledge. Improving the access to quality, precise, reliable and credible information is a main challenge in the medical domain.

At present, there are no barriers to the entry of publication purposes on the WWW. At the same time, there are no standards existing to mandate the validity of information published on the WWW [17]. Health information from the web still lacks of treatment validation. Thus, simply finding information from the web does not provide any guarantee on its correctness and reliability. However, correct health information is of utmost importance because unreliable knowledge and prescription could cause potential dangerous in health, and even in life [18].

Moreover, information available on the WWW is increasing at a staggering rate. The web is estimated to contain over one trillion pages of information currently and is expected to grow by 25 % each year [19]. Due to the ever-increasing volume of information available in the web, once a physician is searching for medical information from the web, he/she will easily become information overloaded [16]. A demonstration has been done by typing in the search query "swine flu" in two advanced search engine, Google Scholar and Medline Plus, which are commonly utilized by physicians. Both of the search results indicate a large number of retrieved articles. Also, the results are not classified into categories, nor arranged according to their relevancy or importance. These make the retrieval process more difficult. Moreover, dissatisfaction with the speed of information retrieval would occur when physicians are lack of sophisticated knowledge and skills in making internet search query. Attempts to manage these issues, there is an emerging need to provide a platform for healthcare professionals to exchange and share knowledge, as well as to facilitate medical information retrieval from the WWW.

Text mining and web mining

Text mining is the process of deriving high-quality information from natural language texts, which are mainly unstructured or semi-structured data [20]. With the rapid growth of Internet use, textual information is pervading the Internet due to the advancement in conversion of paper-based documents into an electronic format. In addition, in the Internet world, text has become the most common mean for formal exchange of information [21]. Thus, web mining is introduced to extract meaningful patterns from massive amount of information on the Internet, through text mining techniques [12]. Web mining is currently gaining attention in different domains due the fast expansion of the WWW, which make Internet become a huge source of knowledge and information. Furthermore, as there are key-phrases that are its semantic metadata to characterize the documents and to produce and overview of the content of the document [22], text mining and web mining can support users with methodologies to compare documents, rank their importance and relevance, and find patterns and trend across different documents [23].

Text mining and web mining have been widely applied in decision support issues in the healthcare domain [24], such as assisting in compiling a medical thesaurus [25], predicting heart attack [26], and delivering reliable and latest information to physicians. For example, Mostafa and Lam [27] used supervised learning to classify with filter medical documents in the area of cell biology; and their study highlighted that a relatively high degree of classification accuracy was achieved by the supervised method. Liu and Chu [28] proposed a knowledge-based query expansion method that exploited the UMLS knowledge source to provide the answers pertinent to certain scenarios that correspond to common tasks performed in medical practice. Elhadad et al. [29] employed a unified user model to create

a tailored summary of relevant documents for physicians by taking advantage of regularities in medical literature text structure and content. Supported by the widely adoption of text mining and web mining in healthcare domain, this study takes into consideration of the supervised learning in classifying the medical documents into particular categories that fulfill the user needs, and hence summarizing the classified documents into short summary, so as to provide reliable medical information to the physicians and achieve the goal of using web information to enhance the quality of decision support of physicians.

Design of the web information retrieval system (WebIRS)

The overall architecture and features of WebIRS are constructed by, IDEF0, a structured system analysis tool, and are presented in Fig. 1. In this proposed system, three modules are involved to control the text processing procedures, including the Retrieval and Preparation Module (RPM), Document Classification Module (DCM) and Text Summarization Module (TSM). RPM helps to retrieve, prepare and preprocess the web information collected from the WWW for further processing and analyzing in the latter modules. In DCM, the Naïve Bayes Classifier [30, 31] is adopted to help identifying and classifying the retrieved articles into different categories. As stated in numerous studies [32–34], generating summary is an efficient method to get control over the information flood. With the concept of summary is to preserve the meaning of the original texts in a shorter but remains informative and readable, TSM intends to generate a short summary for each article for conveying its main idea to the readers. Furthermore, a controlled database, which is a medical library or medical dictionary, is presented to support the entire functioning of the system. There are two purposes for its existence. First, it is used to check up medical jargons that appeared in the retrieved information. Second, it is used to link up medical synonym, for example the word "fever" and "hot", in order to help increase the system's overall performance and accuracy.

A0 Web Information Retrieval System
- A1 Retrieval and Preparation Module
  - A11 Web Crawling
  - A12 Preprocessing
    - A121 HTML Tag Removal
    - A122 Stop Words Removal
    - A123 Words Stemming
  - A13 Keyword Selection and Weight Determination
- A2 Document Classification Module
  - A21 Naïve Bayes Classifier Algorithm
- A3 Text Summarization Module
  - A31 Key Sentences Extraction
  - A32 Key Sentences Arrangement
  - A33 Summary Generation

Fig. 1 IDEF0 architecture of web information retrieval system (WebIRS)

Retrieval and preparation module (RPM)

RPM provides the methodology for the proposed system to retrieve web information from WWW, then process the information into readable and suitable format for later analysis.

Web crawling

Web crawling is the process by which pages from the web are gathered and hence, indexed to support a search engine [35]. It is also referred as a spider for bulk downloading of web pages [36]. The objective of web crawling is to quickly and efficiently gather as many useful web pages as possible, together with the link structure that interconnects them [10]. In WebIRS, the web crawler is responsible for fetching useful web information that could highly match with the health professionals' query need. The web crawler obtains suitable web pages and web contents from time to time, within several web-based medical databases. The time period for the web crawler to gather web information, for example, operating once a week; and also the medical database sources are defined and decided by the physicians. Finally, the fetched materials are stored into the information repository and brought to the next step for further processing. Figure 2 illustrates an overview of the web crawling process.

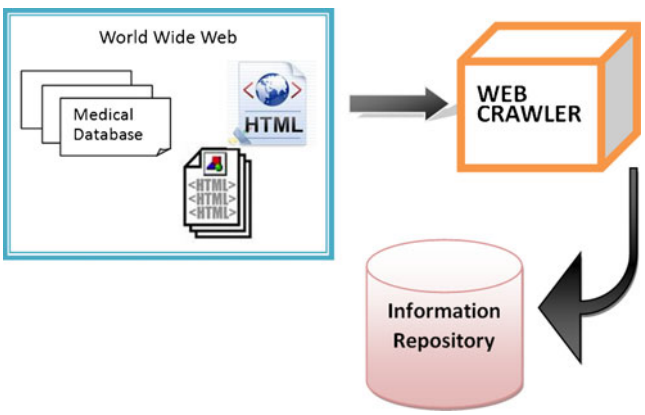

Fig. 2 Web crawling process

Web information preprocessing

After retrieving the useful web information from the web crawling process, several preprocessing steps are conducted to these web documents. All the figures and tables appearing in the web documents are first removed; three preprocessing elements are then applied to improve the performance of text retrieval, classification and summarization [37, 38].

Since the web document displayed in form of HyperText Markup Language (HTML), therefore the disposal of some standard web pages components are conducted. The most common components found are the HTML tags. By interpreting the source code of the web page, all HTML tags (such as <html>, <body>, <p>, <b>, etc.) are being eliminated. The texts wrapped by the HTML tag are taken out, and then generated into plain text file format and being brought to the next preprocessing element. Figure 3 illustrates the procedures done in content extraction.

Upon the completion of HTML tag removal, stop words removal is applied to reduce the noisy information and to improve text processing accuracy [39]. Stop words are words that rarely contribute useful information in terms of document relevance. They are functional words that do not carry any meaning, including articles, prepositions, conjunctions, and some other high frequency words. Examples of these stop words are the, a, in, of, and, it and this. The assumption of stop word removal is that by ignoring the non-informative functional words, assessment of contents of natural language can be facilitated since meaning can be conveyed more clearly, or interpreted more easily [40].

The last preprocessing element is words stemming. As it is necessary to avoid the influence of syntactical features and tenses of the English language when identifying and extracting keywords, words stemming is applied to reduce inflected or derived words to their stem, base or root form [37]. For example, a stemming algorithm for English should stem the words computation, computing, computes, computed, computational, computable, computationally and computers to the root word, compute. It is proven that words stemming has the capability to reduce the redundancy and dimension of the document space representation in an automatic text processing system [32]. Table 1 listed some examples of rules in words stemming.

Keyword selection and weight determination

After preprocessing the web content into a formatted plain text, selecting the keywords for representing a specific class of document is then conducted. According to the literature, numerous methods are proposed to measure the importance of the selected keywords, such as term frequency, inverse document frequency, mutual information, and information gain [Forman, 2003; 41–43]. In this study, the Term Frequency-Inverse Document Frequency (TF-IDF) weight method is employed because it exhibits a superior performance to the others [44]. TF-IDF weight method is a statistical measure used to evaluate and weight the importance of a term or a word to a document within the category collection. With higher TF-IDF weight, the more important a word is towards the document [45]. The TF-IDF weight can be divided into two parts,

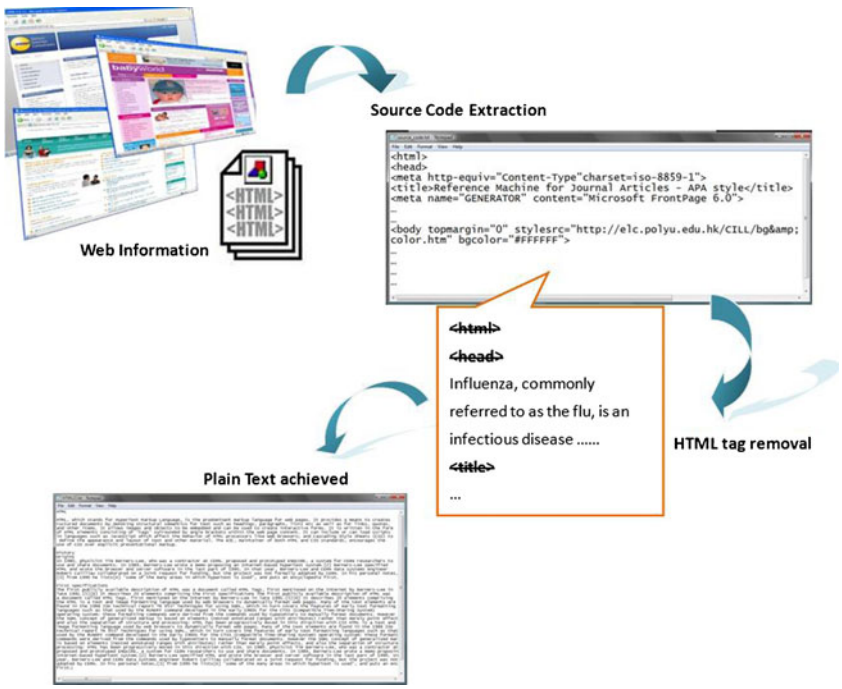

Fig. 3 Procedures of HTML tags removal

Table 1 Examples of rules in words stemming

Rules	Examples
If the word ends in *ed*, remove the *ed*	vaccinated → vaccinat
If the word ends in *ing*, remove the *ing*	coughing → cough
If the word ends in *ly*, remove the *ly*	seriously → serious
If the word ends in *ious*, remove the *ious*	infectious → infect
If the word ends in *es*, remove the *es*	viruses → virus

the Term Frequency (TF) part and the Inverse Document Frequency (IDF) part.

The TF part of the weighting scheme indicates the number of frequency that a word occurs in a document; while the IDF part measures the percentage of all documents within the category collection that contain the given word, thus measures the general importance of the word [46]. The TF-IDF is regarded as a more comprehensive and accurate weight used in text mining because the IDF factor is incorporated to diminish the weight of terms that occur very frequently in the collection, and at the same time to increases the weight of terms that occur rarely. As a result, a high weight in TF-IDF is reached by a high term frequency in the given document, whereas a low document frequency of the term in the whole category collection to avoid biased results [47]. By applying the TF-IDF measurement to all the documents, candidate keywords with term frequency value can be obtained. In order to calculate the TF-IDF weight of each word in a document, first, each term appeared in the preprocessed document is extracted and viewed as a weight vector. The weight vector for document d is denoted as follows:

$$V_d = \left[w_{1,d}, w_{2,d}, w_{3,d}, \ldots, w_{N,d}\right]^T \tag{1}$$

and each term corresponding TF-IDF weight ($w_{i,j}$) is measured with the formula stated as follows:

$$w_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \times log \frac{|D|}{\left|\{d : t_i \in d_j\}\right|} \tag{2}$$

where

$n_{i,j}$	is the number of considered terms (t_i) appeared in the document d_j
$\|D\|$	is the total number of documents in the category collection
$\left\|\{d : t_i \in d_j\}\right\|$	is the number of documents where the term t_i appears.

Document classification module (DCM)

This module initiates a document classifier to categorize the retrieved articles or documents, which are stored in the information repository, into predefined dimensions. In this study, Naïve Bayes Classifier is proposed for document classification. Although there are numerous classifier (such as decision trees, support vector machines, and neural networks), Naïve Bayes Classifier strengths in achieving satisfactory classification accuracy in a relatively short processing time [48–50]. Furthermore, simplicity of the Bayes formula which then requires a relatively small number of training data and shorter training time; and the straightforward calculation and computation required in the building and classification process. Therefore, Naïve Bayes Classifier model is chosen to employ in this study.

Naïve Bayes classifier

Naïve Bayes Classifier is a probabilistic model based on Bayes Theorem to calculate the characteristics of a document using keyword and joint probability of a document category. According to Xhemali et al. [50], Naïve Bayes models are popular in machine learning applications, due to their simplicity in allowing each attribute to contribute towards the final decision equally and independently from the other attributes, and this simplicity is equates to computational efficiency. Naïve Bayes is a supervised leaning model which training is required in the building phase. As stated by McCallum and Nigam [31], the Naïve Bayes Classifier is developed based on the standard Bayes rule defined in Eq. (3).

$$\underset{n}{\text{argmax}}\{P(C_n|d)\} = \frac{P(d|C_n) \times P(C_n)}{P(d)} \tag{3}$$

where

$P(C_n)$	the prior probability of category n
d	the new document to be classified
$P(d\|C_n)$	the conditional probability of the test document, given category n

As $P(w)$ has the same value regardless of the category for which the calculation is carried out, then the category label of w, can be determined by

$$\underset{n}{\text{argmax}}\{P(C_n|d)\} \propto P(d|C_n) \times P(C_n) \tag{4}$$

Classifying stage

With the result of the keyword value performed by TF-IDF weight method, Naïve Bayes Classifier analyzes the text article by extracting the top 100 keywords (with term weight) in each web document. It hence determines the probability of each word being annotated to a particular category. A document is identified and classified to the right category according to the probability of occurrence of certain words in the document that

match with the terms appeared in the list of word occurrence constructed for each category. In other words, the category with the highest probability is assigned to the web document being classified. There may be a situation that the Naïve Bayes Classifier comes up with a result that a particular document can fall into several categories (e.g. Document A is classified in Category A with probability of 0.87 whereas in Category B with probability of 0.55). Thus, in such situation, the classifier arrives at the correct classification as long as the category (i.e. Category A) gives the highest probability value as compared to other categories. All the keywords extracted from the corrected match document are also paired up with the resulting category and this information is recorded in the database to expand the system's classification performance.

Text summarization module

This module intends to develop a method to automatically generate a short summary for an article, which aimed to reduce the time and effort a reader needed to spend on scanning a large number of articles to identify relevant and necessary information [51]. By stemming the document into a short summary of about 100 words, it can help the users to grasp the main idea of the article.

Key sentences extraction

In order to generate a summary that can precisely and correctly deliver the main idea and concepts of an article, it is essential to extract the key sentences from the article. A sentence is regarded as important if it is consisted of as many document's keywords as possible. As mentioned in "Keyword selection and weight determination", TF-IDF weight can help to determine the importance of a term in a document, and hence, help to figure out the document's keywords; therefore, by adding up the TF-IDF weight of each term in a sentence. The sum of TF-IDF weights of a sentence can be calculated to demonstrate the sentence importance within a document. The equation for the calculation of the TF-IDF weight of a sentence, which is denoted as *SentenceScore s*, is stated in Eq. (5).

$$SentenceScore\ s = \sum_{i=1}^{k} w_i \tag{5}$$

where *SentenceScore s* is the total TF-IDF weights of the sentence for term *i* in sentence *s*

Key sentences arrangement

By calculating the sentence score for all sentences in the document, the most important and representing sentence, which is with the highest score, of the document is identified and extracted.

Summary generation

The important sentences are combined to generate the final summary. The sentence with the highest sentence score comes in the first place, following by the sentence with the second high sentence score, and so on. The summary generation stops until it reaches the specified length which is in terms of the number of words, which are 100 words in this proposed system. A demonstrates on the entire summarization procedure is shown in Fig. 4.

Case study–implementation of WebIRS

In order to demonstrate the feasibility of the proposed WebIRS, a leading Hong Kong healthcare organization–Humphrey & Partners Medical Services Limited (HPMS)–was invited to be a case study company. HPMS is one of the largest multi-disciplinary medical services providers in Hong Kong with more than 25 years of practical experience in different disciplines. In this study, ten physicians are invited to have an interview with the purpose of providing advices and consultations throughout the entire system design, development, testing and evaluation processes.

Advises from HPMS

In order to ensure that WebIRS caters the needs of the healthcare domain, seeking physicians' instructions and advices, reviewing and analyzing their users' requirements, and collecting users' feedbacks for system improvement are critical steps in the information system development process [52, 53]. Thus, advisory concerning three areas, including system requirement, interface design, and system performance, are obtained from HPMS.

System requirement

To initiate a system that helps physicians to classify and summarize retrieved web information, first, identification on the number and types of categories that retrieved articles has to be classified into is needed. After consulting from the ten physicians, they suggested classifying the web information into four categories, which are "Disease", "Therapy", "Drug" and "Vaccine".

Upon the completion of the category formation, the information sources for WebIRS to retrieve relevant medical information have to be identified. Four online medical databases, which are most frequently used by physicians in HPMS, including "SCOPUS", "PubMed", "Medline Plus"

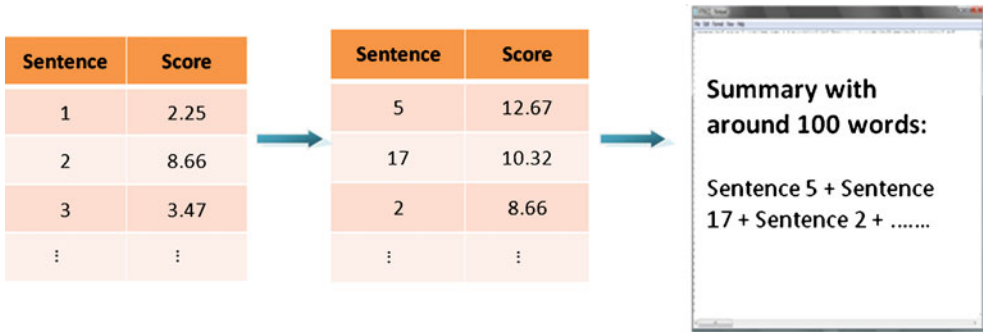

Fig. 4 Summary generation in text summarization module

and "Embase" are being selected. These four web-based medical databases are also regarded as the most popular and reliable information sources within the healthcare domain. The proposed system will then search and find English language articles within these four internationally-recognized medical databases.

Since one of the system's objectives is to visualize the retrieved documents into a short summary for quick review, the ten physicians were therefore consulted to determine the number of words that has to be contained in the generated short summary. It is found that around 100 words would be effective enough to bring about the key ideas and information that could highly represent the entire article. Physicians also suppose that a 100-word summary would be a reasonable and acceptable length for them to go through within their tight schedules.

Interface design

WebIRS is as well acting as a platform to allow users to view the analyzed results processed by the system. In order to construct a user-friendly interface layout to facilitate users' navigation, it is important to collect users' taste and suggestions. Concerning the interface design, it is found that physicians prefer a simple and ease to use. As claimed by the users, most of the physicians do not possess sophisticated computer skills; therefore, a simple, consistent and direct graphical user interface is in favored. Users also demand all text materials to be presented in organized matter with a greater font size.

System performance

After employing WebIRS to assist in the information searching process, users' feedbacks concerning the system performance are gathered. The detailed system evaluation and suggested areas for improvement will be discussed in "Performance evaluation and discussion".

System design

With the advices of HPMS, WebIRS is designed and developed to cater the physicians' interests. As for the proposed system, a user platform named MedicPedia is developed as the system user interface. MedicPedia's features and interface with corresponding descriptions are shown in Table 2.

Scenario demonstration

This section demonstrates the proposed web information retrieval system that is developed for the physicians taking into consideration the advices of HPMS and system requirement presented in the previous sections. The first step of the system application is user authentication. Before accessing the MediaPedia for medical information retrieval, registered users have to complete the login step by entering the login ID and password into the login page (Fig. 5). Furthermore, logos of the four medical databases, which are the information sources of the system, are shown in the login page to realize where the sources come from.

After a successful login, a welcome page is returned and the last login information and the list of suggested readings are provided (Fig. 6). User can simply view the date and time of their previous login, as well as the last article they have read during the previous login. Also, five suggested articles are given to the users according to their previous reading preference. Take the case of Dr. Lo, a specialist in Upper Respiratory Tract Infection (URTI) working in HPMS, as an example. The system has a record of his previous readings that indicates he is most interested in topics and articles with the keyword "respiratory". As a result, when Dr. Lo login MediPedia, the system will automatically generate five articles with the keyword "respiratory" as suggested reading materials. And these five articles would be those Dr. Lo have not read before, and are the newest articles that the system currently retrieved. Similar to all articles stored in the system, a short summary about the article is provided.

To start with the information retrieval process, users can adopt two means. The first one is by selecting the interested category, and the second one is search by search query. Regarding the search by category, for example, as the user is interested to view articles about the use and prescription of a specific drug, or wanted to know more about the side effects of a particular drug. He/she can simply click into the

Table 2 MediPedia's features and interface with corresponding descriptions

MedicPedia features	Description
User authentication	In order to utilize the functions provided by MedicPedia, a user has to register for a login account first. A login page is designed for the authentication of a registered user identity.
Last login information	MedicPedia records each users' login history and the readings that they read. During users' next login, the previous login date and time, with the last article being read will be shown in the welcome page as a reminder.
Suggested readings	To provide a customized platform to enhance users' satisfaction, MedicPedia analyzes the users' reading record to identify the users' most frequently read topic or category; and thus, provide a list of five suggested articles to in every login.
Navigation panel	The navigation panel directs the users to use the platform by either through selection of the interested category, or by directly typing in the search query into the search box.
Document classification result	MedicPedia sorts and displays the classified articles from the highest to the lowest according to their relevance towards the particular category. And a total of number of five results is shown per page.
Text summarization result	The short summary generated for each article would be displayed right below each article's title. Thus, when a user finds a particular article title interesting or useful, he/she can directly refer to the summary below to know more what the article is about.

category "Drug" to obtain articles which are related to this topic. As the delivered results have undergone the document classification process, they are regarded as highly relevant to the category "Drug". The percentage of relevancy of each article towards the respective category is indicated before the article title for users' reference. The screen capture of the classification result with indication of the relevancy percentage is shown in Fig. 7.

When the user, who is interested in information about drugs, found the article "Beta Blockers May Slow Spread of Breast Cancer" has a relatively high relevancy towards the category, and supposes it could be a useful and worth-reading one. To know more about the article content and to verify whether it is a relevant material before going deep into the entire article, he/she can select "View FULL summary" to view the generated 100-word article summary. The screen capture of "View FULL summary" selection is shown in Fig. 8.

After reading the summary of the article, the user finds the article interesting, and then he/she can click the article title which is indicated in blue color to view the entire article. The full version of the desired article is then presented in the lower part of the browser.

The second way to search for information in MedicPedia is to input a search query into the search box at the navigation panel. Take another case as an example, the user now wanted to search for information regarding "swine flu". So, he inputs the word "swine flu" into the search box, and then a list of classified articles is retrieved. A screen capture of the search result is shown in Fig. 9.

Unlink other current search engine, MedicPedia classifies the research results into categories. As a result, when the user initiates a search query "swine flu", the retrieved articles about swine flu would be classified into four categories, which are "Disease", "Therapy", "Drug" and "Vaccine". By providing a classified result, it can help user to locate his/her

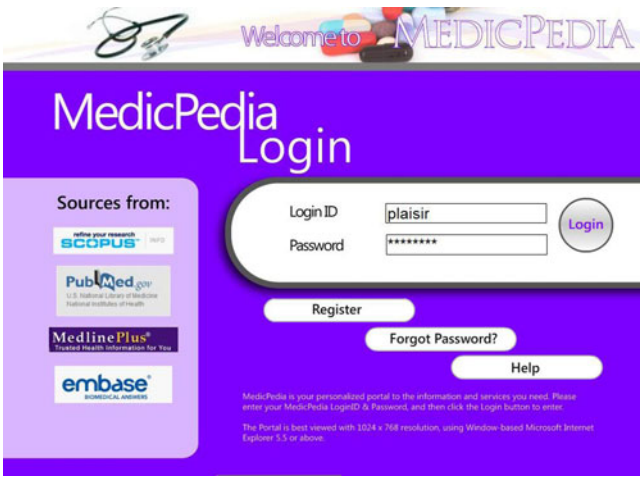

Fig. 5 Login page of MedicPedia

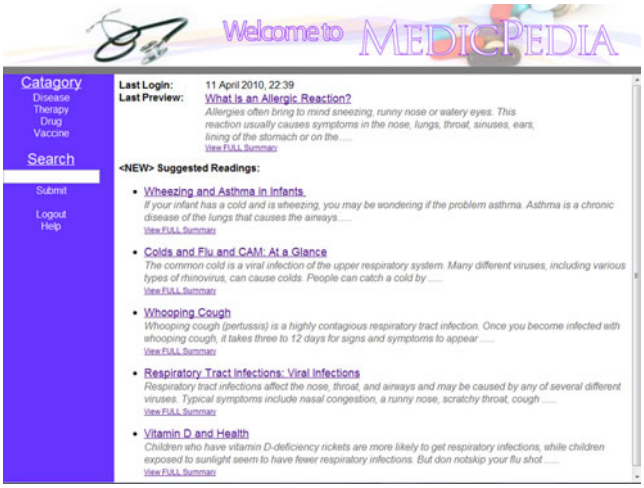

Fig. 6 Welcome page of MedicPedia

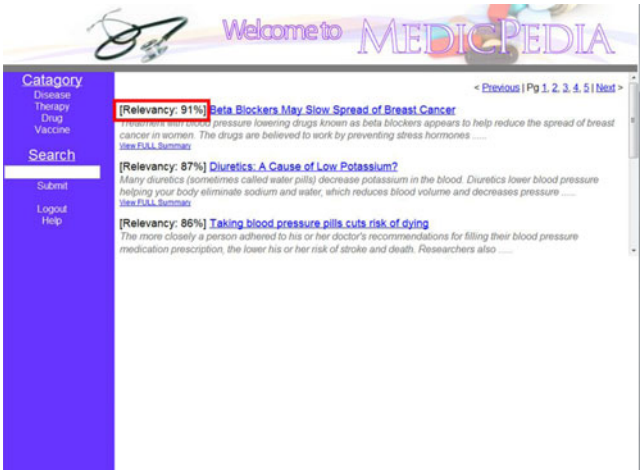

Fig. 7 Classification result

Fig. 9 Search result by search query

interested articles in a quicker and easier manner. The screen captures of the classified results for swine flu with category indication are shown in Fig. 10.

Whereas similar to the article retrieval process mentioned in the previous part, search by category, when the user find the article "2009 H1N1 Flu (Swine Flu) Symptoms" suits his/her needs and interests, he/she can select to view the entire summary of the article, or click into the article title to start reading the desired article. After the searching of all necessary information and articles, the user has to click logout on the navigation panel to exit the system. The system will then return to the login page of the system after successful logout.

Performance evaluation and discussion

Keyword accuracy

The classification performance of the system is highly dependent on the keywords that the system could extract from an article. It is because to determine which category an article would fall into, the keywords being extracted are the major decision criteria. With more keywords that could match with the category keywords list, an article is more likely to be classified into that category. The word accuracy result is introduced to evaluate the system's effectiveness to extract relevant and accurate keywords from an article to attain better classification performance.

Fig. 8 "View FULL summary" selection

In the proposed system, six keywords will be extracted from each article. These six keywords are being selected according to their TF-IDF weight. With higher TF-IDF weight, the more important a word is towards an article; therefore, the six words with the highest TF-IDF weight would be selected as the keywords of an article. It is also find that in every international journal and paper, a set of article keywords that are defined by the article's author can be found as the article information. As a result, in the evaluation the keyword accuracy, the two set of keywords, the one defined by the author and the one extracted by the system, will be used to compare against one another. By comparing the two set of keywords, the number of matched keywords between the two keyword set can be generated. The system's ability to extract accurate keywords form an article can then be measured by Eq. (6). Higher percentage of accuracy means the system can automatically extract keywords that highly match with the article's original keyword set; and therefore, the system is pursuing satisfactory capability in keywords extraction.

$$\text{Keyword Accuracy} = \frac{\text{Number of matched keywords}}{\text{Number of article's keywords}} \times 100\% \qquad (6)$$

To evaluate the proposed system's keywords extraction ability, 200 thesis which collects from the database of PolyU Institutional Repository (http://repository.lib.polyu.edu.hk) are selected for the assessment. The number of article's keywords, number of retrieved keywords, number of

Fig. 10 Classified search result with different categories

matched keyword, as well as the calculation of the accuracy percentage are listed in Table 3.

As shown in Table 3, it is found that more than half of the articles can bring about a keyword accuracy of over 70 %, and nearly all articles can achieve above 50 % keyword accuracy. Therefore, it is concluded that the proposed system is able to extract accurate keywords from an article at a reasonable performance level.

Users' feedback

Feedbacks from the ten physicians in HPMS, who have used the system in 1-month period to assist in the information retrieval process, are collected. Face-to-face questionnaires are conducted with them to evaluate the system's performance.

Table 3 Keywords information of the 200 articles

Article #	No. of article's keywords	No. of retrieved keywords	No. of matched keywords	% of accuracy
1	4	6	3	75 %
2	5	6	3	60 %
3	6	6	5	83 %
4	6	6	4	67 %
5	4	6	2	50 %
6	5	6	2	40 %
7	6	6	4	67 %
8	4	6	3	75 %
.	.	.	.	.
.	.	.	.	.
199	5	6	4	80 %
200	5	6	3	60 %

During the administration, they were asked about the differences in their web searching process before and after the use of the proposed system. It is found that there are three significant differences, namely the number of articles being read, the time spent on searching, and the relevancy of the article. The results of these three differences are summarized in Table 4.

It is realized that without any assistant from the web information retrieval system, physicians have to search through the general search engine, and read on an average of 6 articles and spend around 20 min, before they can sort out a relevant article that suit their needs. After using the system, physicians reported that the average number of articles being read and the average time needed to retrieve an useful article have significantly reduced to 2 articles and about 5 min respectively. Yet, a drop from 98 % to 85 % is found in the users' estimated relevancy of the retrieved articles due to the reason that, in the past, physicians need to go through the text content of the articles one after another before they come up with an suitable and useful one. Thus, the article is selected after the physician has spent great effort to read through majority of its details. But after using WebIRS, the physicians would consider the article to be relevant mainly based on the summary of the article, but without prior investigation in the articles' content. As a

Table 4 Summary of users' feedback

	Before	After
Number of articles read	6	2
Time spent on searching	20 min	5 min
Relevancy of articles	98 %	85 %

result, the users' estimated relevancy of a retrieved article may be diminished.

Although there is a fall in the relevancy of the retrieved articles, physicians claimed that the system performance is still satisfactory and they would continue to adopt it. It is because the reduction in searching time and the number of articles needed to be read outweigh the drop in articles' relevancy. Therefore, through users' feedback, the system is regarded as a useful one with excellent performance.

Overall benefits to the healthcare professionals

Reduction in web search time

Since the system is trained by the classified article results provided by the physicians, the retrieval result will be much better compared with the traditional method, which is by random keyword searching manually. As mentioned previously, physicians require less time in retrieving what they want when they are using the proposed system. Furthermore, the document they need to read and review reduced as the proposed system will filter all the irrelevant documents as well as to summarize the document in around 100 words for better reading. As a result, the web search time significantly falls when the proposed system is used.

Effective summarization result

The summarization capability helps physicians in understanding the document in a short period of time. The 100-words summary is rather useful as it gathers all the important information and hence displays to the physicians. Therefore, the summary enables the users to get vital concepts or keywords about the articles when compared with the traditional approach (i.e. only display the first page of the paper).

Reliable and credible information sources

Since all the documents are retrieved under the data sources, which are the four internationally recognized web-based medical databases, provided by the physicians; therefore, all the retrieved information are reliable and credible. When compared with the traditional approach where physicians need to perform random searching via the Internet, the creditability of information sources is guaranteed.

Efficient web-based information retrieval platform

The web-based information retrieval platform provides the latest medical information to physicians in a timely manner. The web-based capability ensures users to access to the information at anytime and anywhere. As mentioned previously, online information has become a main source to help improve the quality of services and decision making of physicians; the proposed system is demonstrated as an efficient and promising tool in web information retrieval system in the case study implementation.

Conclusions and further research directions

In the current information technology era, the WWW is gaining popularity and importance. Internet has become one of the most common and convenient sources for obtaining information and knowledge. This phenomenon also applies to the healthcare domain, where medical professionals are always keen to obtain the latest and accurate information. Nevertheless, due to the constraints of the current web information retrieval process and the limitations of web search engine, physicians are facing the problems of information overload, information credibility and information timeliness when they are initiating a search query via internet. To address and improve the current situation, this paper proposes a WebIRS for the health professionals to navigate the system processed results.

With the aim to assist and facilitate the web information retrieval process in the medical field, the proposed WebIRS arrange the retrieved medical documents according to highest degree of correlation of categories to the lowest one. Thus, physicians can easily identify their information needs and obtain the relevant information that suits their needs and interests. Furthermore, with the short summary provided for each article, physicians no longer need to go through the entire article to scan for its main thought before justifying whether the information is necessary or not. By simply reading the 100-words summary, physicians may grasp the core message suggested by the article. As a result, these can help medical professionals to achieve a more accurate, effective and efficient information retrieval process.

The proposed information retrieval approach has been validated in a medical center. The satisfactory results demonstrate the potential for adoption of this method in various medical organizations. However, there is still room for further development. Further research will consider classification categories and information sources into the system to determine a more comprehensive web information retrieval system. Furthermore, the consideration of retrieving articles with different languages can be tested. As different languages have different syntactical features, time and effort are needed to research and analyze on their different information representation methods. Thus, articles of different languages could be processed by the system to boost its comprehensiveness.

Acknowledgment Acknowledgement is given to Dr. Peter Lo, Dr. Francis Liu, Dr. C.W. Lo and Miss Maggie Poon for their guidance on issues in clinical coding and medical knowledge in general. The authors would also like to express their sincere thanks to the Research Committee of the Hong Kong Polytechnic University for providing the financial support for this research work.

Conflict of interest statement There are no potential conflicts of interest in this paper.

References

1. Housel, T., and Bell, A. A., *Measuring and managing knowledge*. McGraw Hill, Irwin, 2001.
2. Bell, G. B., and Sethi, A., Matching records in a national medical patient index. *Commun. ACM* 44(9):83–88, 2001.
3. Ybarra, M. L., and Suman, M., Help seeking behavior and the Internet: a national survey. *Int. J. Med. Inform.* 75(1):29–41, 2006.
4. Gilmour, J. A., Scott, S. D., and Huntington, N., Nurses and Internet health information: a questionnaire survey. *J. Adv. Nurs.* 61(1):19–28, 2008.
5. McHugh, S. M., Corrigan, M., Morney, N., Sheikh, A., Lehane, E., and Hill, A. D. K., A quantitative assessment of changing trends in Internet usage for cancer information. *World J. Surg.* 35(2):253–257, 2011.
6. Rogers, S. N., Rozek, A., Aleyaasin, N., Promod, P., and Lowe, D., Internet use among head and neck cancer survivors in the North West of England. *Br. J. Oral Maxillofac. Surg.* 50(3):208–214, 2012.
7. Holzinger, A., Geierhofer, R., Modritscher, F., and Tatzl, R., Semantic information in medical information systems: utilization of text mining techniques to analyze medical diagnoses. *J. Univ. Comput. Sci.* 14(22):3781–3795, 2008.
8. Casebeer, L., Bennett, N., and Kristofco, R., Physician Internet medical information seeking and on-line continuing education use patterns. *J. Contin. Educ. Health Prof.* 22:33–42, 2002.
9. Jennings, N. R., and Wooldridge, M. J., *Agent technology: foundations, applications, and markets*. Springer, Berlin, 1998.
10. Manning, C. D., Raghavan, P., and Schütze, H., *Introduction to information retrieval*. Cambridge University Press, U.K., 2008.
11. Vakali, A., and Pallis, G., *Web data management practices: emerging techniques and technologies*. Idea Group Publishing, U.S.A., 2007.
12. Velasquez, J. D., and Palade, V., *Adaptive web sites: a knowledge extraction from web data approach*. Ios Press, Netherlands, 2008.
13. Tao, X., Li, Y., and Zhong, N., A knowledge-based model using ontologies for personalized web information gathering. *Web Intell. Agent Syst.* 8(3):235–254, 2010.
14. Kalichman, S. C., Weinhardt, L., and Benotsch, E., Internet access and internet use for health information among people living with HIV-AIDS. *Patient Educ. Couns.* 46(2):109–116, 2002.
15. Risch, N. A., Kwon, H. T., and Scarbrough, W., Minority primary care physicians' knowledge, attitudes, and practices on eye health and preferred sources of information. *J. Natl. Med. Assoc.* 101(12):1247–1253, 2009.
16. Walczak, S., A multiagent architecture for developing medical information retrieval agents. *J. Med. Syst.* 27(5):479–498, 2003.
17. Craan, F., and Oleske, D. M., Medical information and the Internet: do you know what you are getting? *J. Med. Syst.* 26(6):511–518, 2002.
18. Ku, Y., Chiu, C., and Liou, B. H., "Applying text mining to assist people who inquire HIV/AIDS information from Internet". In: *Proceedings ISI 2008 Workshops*. pp. 440–448, 2008.
19. Szulencki, P., "Number of pages on Internet according to Google", available at: http://www.seoblogr.com/google/number-of-pages-on-internet-according-to-google (accessed 15 April 2011), 2008.
20. Antonio do Prado, H., and Ferneda, E., *Emerging technologies of text mining: techniques and applications*. Information Science Reference, U.S.A, 2008.
21. Berry, M. W., *Survey of text mining: clustering, classification and retrieval*. Springer, New York, 2004.
22. Song, M., and Wu, Y. F., *Handbook of research on text and web mining technologies*. Information Science Reference, U.S.A., 2009.
23. Han, J., and Kamber, M., *Data mining: concept and techniques*. Morgan Kaufmann, San Francisco, 2006.
24. Ting, S. L., Shum, C. C., Kwok, S. K., Tsang, A. H. C., and Lee, W. B., Data mining in biomedicine: current applications and further directions for research. *J. Softw. Eng. Appl.* 2(3):150–159, 2009.
25. Lu, W. H., Lin, R. S., Chan, Y. C., and Chen, K. H., Using Web resources to construct multilingual medical thesaurus for cross-language medical information retrieval. *Decis. Support. Syst.* 45(3):585–595, 2008.
26. Patil, S. B., and Kumaraswamy, Y. S., Intelligent and effective heart attack prediction system using data mining and artificial neural network. *Eur. J. Sci. Res.* 31(4):642–656, 2009.
27. Mostafa, J., and Lam, W., Automatic classification using supervised learning in a medical document filtering application. *Inf. Process. Manag.* 36(3):415–444, 2000.
28. Liu, Z., and Chu, W. W., Knowledge-based query expansion to support scenario-specific retrieval of medical free text. *Inf. Retr.* 10(2):173–202, 2007.
29. Elhadad, N., Kan, M. Y., Klavans, J. L., and McKeown, K. R., Customization in a unified framework for summarizing medical literature. *Artif. Intell. Med.* 33(2):179–198, 2005.
30. Lewis, D. D., "Naive Bayes at 40: the independence assumption in information retrieval". In: *Proceedings of the 10th European Conference on Machine Learning*. pp. 4–15, 1998.
31. McCallum, A., and Nigam, K., "A comparison of event models for Naive Bayes text classification". In: *Proceedings of AAAI-98 Workshop Learning for Text Categorization*. 1998.
32. Zhan, J. M., Loh, H. T., and Liu, Y., Gather customer concerns from online product reviews—a text summarization approach. *Expert Syst. Appl.* 36(2):2107–2115, 2009.
33. Lloret, E., Llorens, H., Moreda, P., Saquete, E., and Palomar, M., Text summarization contribution to semantic question answering: new approaches for finding answers on the web. *Int. J. Intell. Syst.* 26(12):1125–1152, 2011.
34. Fan, W., Wallace, L., Rich, S., and Zhang, Z., Tapping the power of text mining. *Commun. ACM* 49(9):76–82, 2006.
35. Cothey, V., Web-crawling reliability. *J. Am. Soc. Inf. Sci. Technol.* 55(14):1228–1238, 2004.
36. Olston, C., and Najork, M., *Web crawling*. Now Publishers Inc, Hanover, M.A., 2010.
37. Salton, G., *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley, Reading, M.A., 1989.
38. Yang, Y., and Chute, C. G., An example-based mapping method for text categorization and retrieval. *ACM Trans. Inf. Syst.* 12(3):252–277, 1994.
39. Chakrabarti, S., Roy, S., and Soundalgekar, M. V., Fast and accurate text classification via multiple linear discriminant projection. *Int. J. Very Large Data Bases* 12(2):170–185, 2003.
40. Patwardhan, S., and Pedersen, T., *Using WordNet-based context vectors to estimate the semantic relatedness of concepts*. National Science Foundation Faculty, U.S.A., 2006.
41. Chen, J., Huang, H., Tian, S., and Qu, Y., Feature selection for text classification with Naïve Bayes. *Expert Syst. Appl.* 36(3):5432–5435, 2009.

42. Mladeni, D., and Grobelnik, M., Feature selection on hierarchy of web documents. *Decis. Support. Syst.* 35(1):45–87, 2003.
43. Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., and Wang, Z., A novel feature selection algorithm for text categorization. *Expert Syst. Appl.* 33(1):1–5, 2007.
44. Lan, M., Sung, S. Y., Low, H. B., and Tan, C. L., "A comparative study on term weighting schemes for text categorization". In: *Proceedings of International Joint Conference on Neural Networks (IJCNN-05)*. Vol. 1, pp. 546–551, 2005.
45. Aizawa, A., An information-theoretic perspective of tf-idf measures. *Inf. Process. Manag.* 39(1):45–65, 2003.
46. Radev, D. R., Jing, H. Y., and Tam, D., Centroid-based summarization of multiple documents. *Inf. Process. Manag.* 40(6):919–938, 2004.
47. Wu, H. C., Luk, Y. P., and Wong, K. F., Interpreting TF-IDF term weights as making relevance decisions. *ACM Trans. Inf. Syst.* 26(3):13–37, 2008.
48. Isa, D., Kallimani, V. P., and Lee, L. H., Using the self organizing map for clustering of text documents. *Expert Syst. Appl.* 36(5):9584–9591, 2009.
49. Lin, S. S., A document classification and retrieval system for R&D in semiconductor industry—a hybrid approach. *Expert Syst. Appl.* 36(3):4753–4764, 2009. Part 1.
50. Xhemali, D., Hinde, C. J., and Stone, R. G., Naive Bayes vs. decision trees vs. neural networks in the classification of training web pages. *Int. J. Comput. Sci. Issues* 4(1):16–23, 2009.
51. Ou, S., Khoo, S. G., and Goh, D. H., Automatic multidocument summarization of research abstracts: design and user evaluation. *J. Am. Soc. Inf. Sci. Technol.* 58(10): 1419–1435, 2007.
52. Jacobsen, I., Booch, G., and Rumbaugh, J., *The unified software development process*. Addison Wesley, Boston, M.A., 1999.
53. Ting, J. S. L., Kwok, S. K., Tsang, A. H. C., Lee, W. B., and Yee, K. F., "Experiences sharing of implementing template-based electronic medical record system (TEMRS) in a Hong Kong medical organization". *J. Med. Syst.* 35(6):1605–1615, 2011.