Research and Innovative Design of Search Engine for Banking Industry Decision-makers

Huaihai Hui
School of Economics and Management,
Chinese Academy of Sciences,
Beijing, P.R.C.
huihuaihai@ucas.ac.cn

Des McLernon
School of Electronic and Electrical Engineering,
University of Leeds,
Leeds, LS2 9JT United Kingdom.
d.c.mclernon@leeds.ac.uk

Ali Zaidi
School of Electronic and Electrical Engineering,
University of Leeds,
Leeds, LS2 9JT, United Kingdom.
s.a.zaidi@leeds.ac.uk

ABSTRACT

In order to solve the problem that General Search Engines (GSEs) involve a wide range of industries, the amount of information obtained through the search is large, information is disorderly arranged, queries are inaccurate, and it is not sufficiently professional. Based on the actual needs of the banking industry, this paper designs and develops an innovative Search Engine for Banking Industry Decision-makers (SEfBIDm). This paper presents the needs analysis, overall functional design, overall framework and workflow design of SEfBIDm. SEfBIDm can provide many functions such as a banking knowledge database, image search and analysis report. This article only gives the implementation method and workflow of typical functions of web search. SEfBIDm was deployed, tested, and operated by 69 branch decision-making agencies at the world's 10th largest bank. Decision-makers from these decision-making bodies believe that SEfBIDm is rooted in the banking industry and that it is supported by banking industry knowledge and experts, which are not available in GSEs. The information obtained from SEfBIDm has three distinct characteristics: the support of comprehensive information before decision making; the timeliness of feedback information during the execution of decisions; and the very accurate evaluation information obtained when the decision is executed.

CCS Concepts

• Information systems → Information retrieval → Search engine architectures and scalability → Search engine indexing.

Keywords

Decision-makers, Banking Industry, Search Engine.

1. INTRODUCTION

With the basic popularity of the traditional Internet and the rapid development of the mobile Internet, the growth of network data has begun to show a geometric series growth trend. Human society has entered a real information age. As an effective tool for obtaining information, search engines play a key role in the development of the information society. However, a GSEs has many deficiencies in itself while achieving great success [1]. First of all, a GSEs involves all fields and industries, so it is not specialized in certain industry-specific queries; secondly, a GSEs is based on the full-text search of keyword matching or category browsing based on the topic of the website. It may ignore the content that users really need to query. This can lead to erroneous search results or missing key search content. Therefore, how to design and implement a search engine that targets a specific user in a specific industry has become a topic that needs to be solved [2]. Based on the actual needs of the banking industry, this paper designs and develops an innovative Search Engine for Banking Industry Decision-makers (SEfBIDm).

2. NEEDS ANALYSIS

Decision-makers refers to the subject who makes decisions based on social, political, economic, cultural, and psychological factors. The decision-makers can be an individual, such as the general manager of a company, or a group, such as a company's board of directors. Decision-makers need a high level of competence. For example, they need to discover the fundamental characteristics of problems, need to identify decision opportunities effectively, need to deal with complexity, and need to accurately assess the resources needed to implement decisions [3]. In particular, they need comprehensive, timely and accurate information to deal with the various uncertainties of decision-making.

For the specific group of decision-makers in the banking industry, their ability requirements are higher than those of other decision makers. This is because with the rapid development of social information, more and more customer data and business data have been accumulated by banks, and the industry has become increasingly competitive. In addition to maintaining its traditional business and services, banks also need to use Internet and banking industry data resources to enhance their competitiveness [4].

For the decision-makers in the banking industry, they also need to have comprehensive, timely and accurate banking information. They need comprehensive information support before making decisions. They need to get timely feedback during the execution of the decision. At the same time, they need to obtain accurate information to evaluate their decisions after the decision has been executed. They do not need the massive, unordered, unstructured information that a GSEs provide.

Therefore, specialized search engines for banking industry decision makers need to search for specific information content in the professional websites of the Internet and banking industry. So, we require a more advanced GSEs.

This kind of professional search engine needs to simplify, integrate, and direct the information in the web pages of the banking industry, and then extract the required data in the sub-fields. The extracted information is processed and returned to the user in a specific form. This engine requires sophisticated processing of banking industry information. It needs to solve the problem that the GSEs obtains too much information, disorderly information, and inaccurate query contents.

3. OVERALL FUNCTIONAL DESIGN

The main functions of SEfBIDm are information capture functions, information analysis functions, information indexing and retrieval functions, and the establishment of a banking industry knowledge database.

3.1 Information Capture Functions

SEfBIDm has a smaller search range and high search efficiency. SEfBIDm uses the seed Webpage from the URL database as the starting point for the search. The content of the SEfBIDm index is limited to the bank subject or a related field, so the captured information is more inclined to structured data and metadata. SEfBIDm's information crawling function is performed by web crawlers. SEfBIDm's crawlers include image crawlers and web crawlers.

The image crawler is responsible for crawling the corresponding pictures on the URLs of all pictures stored in the img table in the database. It includes major picture formats such as jpg, jpeg, gif, and png. The web crawler is responsible for crawling web pages saved in the URL table in all databases.

The image crawler consists of two parts: thread control and crawling thread. The thread control module is responsible for reading the configuration file of the program, connecting to the database and obtaining the URL of the image that is not crawled, and then assigning it to each crawling thread. The crawling thread crawls the pictures one by one according to their assigned URL. When the crawling is over, the URL allocation function of the control module is invoked to allocate a new URL for itself. The image needs to be compressed and stored as a thumbnail when searching. When the control thread finds no crawlable URL in the database, the process sleeps.

The basic process of the web crawler is described as follows:

1) Set up websites and keywords that need to be crawled;

2) Select the URL and keywords to be crawled;

3) Scheduling crawler threads;

4) The URL is assigned to the crawler thread;

5) Determine whether the URL is assigned. If the allocation is completed, go back to step 2 and repeat. Otherwise, go to the next step;

6) The crawler crawls the URL and analyses the crawled page. At the same time, change the relative path to an absolute path;

7) Output pages to the cache;

8) Determine whether the crawling is completed. If it is not completed, repeat steps 7 and 8, otherwise skip to step 6.

3.2 Information Analysis Functions

The task of the SEfBIDm analyzer includes the relevance analysis of web pages, the evaluation of web pages, the filtering of web pages, the analysis of texts, word segmentation and word frequency statistics.

1) Carry out the relevance analysis of the webpage to determine whether it is a webpage related to the banking industry;

2) Give the evaluation of the link and determine whether it is a valid link, so as to determine whether the searcher should grab the link in the next step;

3) The filtering of web pages automatically filters out unwanted advertisements, pop-up windows, etc., leaving the main part for subsequent processing;

4) Document analysis and word segmentation. The topic content is analyzed from the filtered web page. The frequency and number of occurrences of general words, professional words, and classification feature data are counted by word frequency.

3.3 Indexing and Retrieval Functions

SEfBIDm uses a special data storage format that relies on indexing mechanisms and algorithms. This technology meets the requirements of fast implementation, easy maintenance, fast retrieval, and low space requirements of the index model. This kind of indexer is a completely different structure from the normal database. It outputs the index table. It uses the inverted list to find the corresponding document based on the index item.

SEfBIDm performs semantic analysis based on the user's keywords, queries the documents in the index database, sorts the results to be output, and implements a relevance feedback mechanism for decision-makers in the banking industry. Information retrieval technology based on natural language understanding is a development trend of dedicated search engines [5]. In addition to the traditional search engine, it provides functions such as traditional quick search and relevance ranking. It also provides semantic understanding of the content, intelligent information filtering and pushing, user role records, automatic recognition of user interests and other functions.

3.4 Banking Industry Knowledge Database

The establishment of the banking industry knowledge database is achieved through two methods: One is to organize data on its own, and the other is to search for data already established on the Internet. SEfBIDm uses directional search technology to complete data acquisition and database establishment. There are two modes for searching the banking industry knowledge database. The first is "search terms", which lists the entry abstracts, sources, and snapshots of the webpages related to the user's search; the second is "entering entries," giving detailed explanations of specific entries. No matter which mode is used, users can get suggestions for the closest terms. Users can obtain accurate banking industry knowledge from knowledge databases by entering search terms.

4. FRAMEWORK AND WORKFLOW

The key technology of SEfBIDm is how to obtain effective banking industry information from the Internet. This requires improvement of the GSEs. First of all, there is a need to combine proprietary search engines and strategies in the banking industry. Secondly, it needs to effectively capture information from web pages and store it structurally in a database. Thirdly, it is necessary to index the information stored in the database and establish an effective indexing mechanism. Then, it needs an effective search of the information in the database. This not only needs to fuzzy match the keyword information input by the user, but also needs semantic judgment on the keyword information to search for information that is closer to the user's needs, so as to improve the efficiency of information search. Finally, it needs to build a data model on a database that stores information, and then perform On-Line Analytical Processing (OLAP) analysis to mine potential information [6]. The overall framework of SEfBIDm is shown in Figure 1.

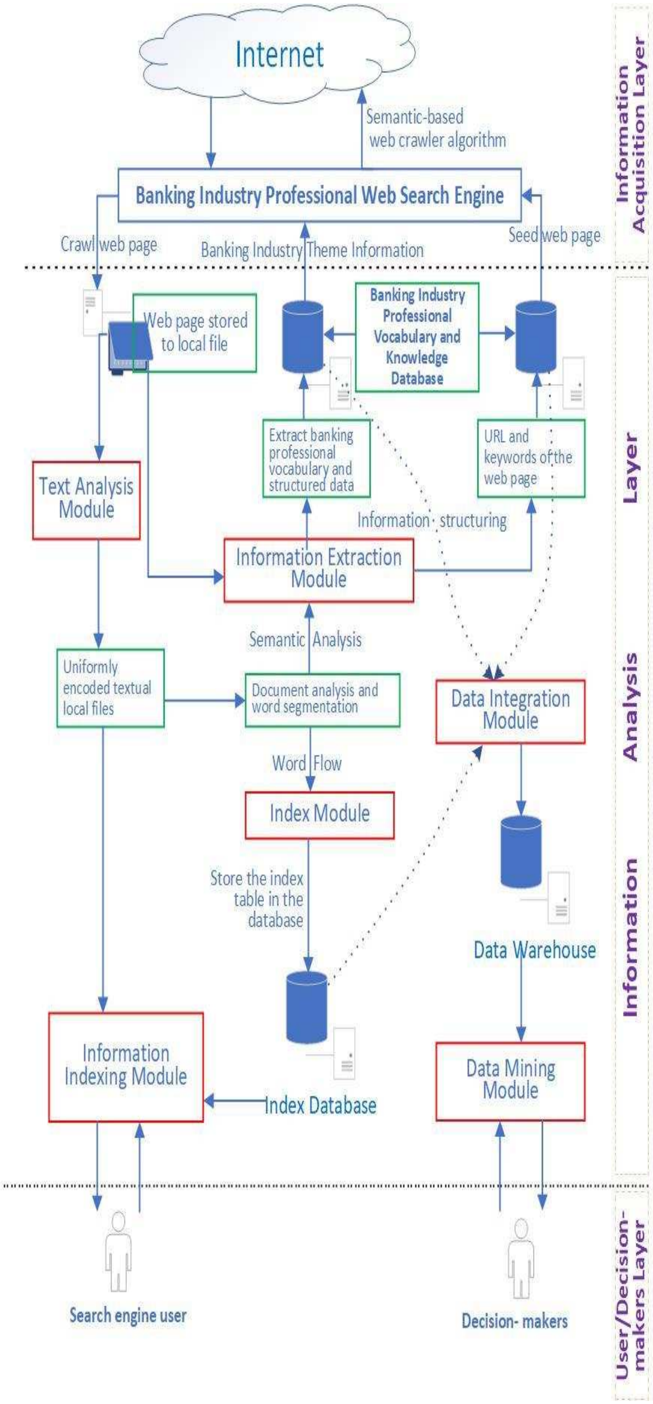
Internet
Semantic-based web crawler algorithm
Banking Industry Professional Web Search Engine
Crawl web page
Banking Industry Theme Information
Seed web page
Information Acquisition Layer
Web page stored to local file
Banking Industry Professional Vocabulary and Knowledge Database
Extract banking professional vocabulary and structured data
URL and keywords of the web page
Text Analysis Module
Information structuring
Information Extraction Module
Semantic Analysis
Uniformly encoded textual local files
Document analysis and word segmentation
Data Integration Module
Word Flow
Index Module
Store the index table in the database
Data Warehouse
Information Indexing Module
Index Database
Data Mining Module
Information Analysis Layer
Search engine user
Decision- makers
User/Decision-makers Layer

Figure 1: The overall framework of SEfBIDm

The SEfBIDm function module includes an information acquisition layer, an information analysis layer, and a user (decision-makers) layer. At the information acquisition layer, SEfBIDm uses a web crawler of a specific algorithm to actively search and crawl web pages related to specified information from a limited range of Internet and banking industry internal web pages. At the information analysis layer, the web page information is processed, and the structure of the captured web page is analyzed. URLs and web page information are stored in the URL database. The text information in the web page is analyzed lexically and an index table corresponding to it is established. At the same time, the information obtained from the web pages is also structured and stored in the database. The function of the user (decision-makers) layer includes two aspects: On the one hand, it implements the user's search for information. SEfBIDm performs semantic analysis on the keywords entered by the user, and then retrieves the information the user needs through the search module and returns it to the user through the web page. On the other hand, it provides users with valuable potential information. It builds a data model of the information in the database and performs OLAP analysis on the data model to analyze the effective information and then pushes it to the bank's decision makers.

The functional modules of SEfBIDm include: searcher, filter, analyzer, indexer, and retriever. The SEfBIDm workflow is shown in Figure 2 below.

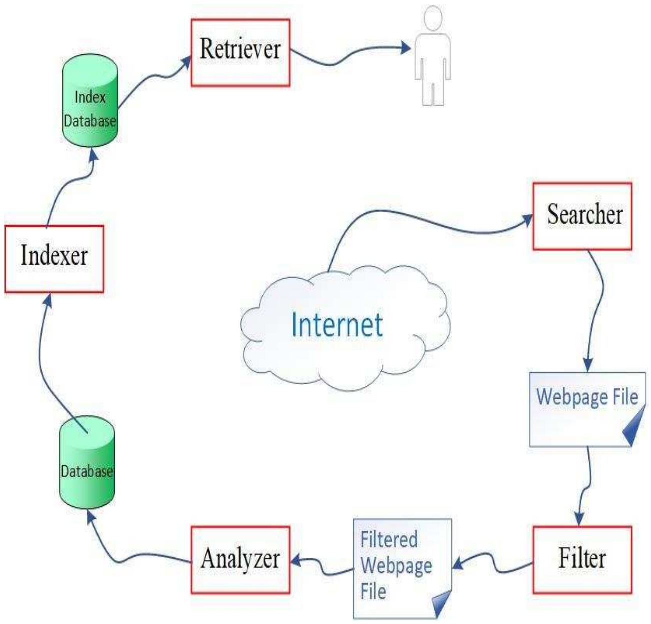

Figure 2: The SEfBIDm workflow

5. WEB SEARCH DESIGN

Web search uses a method of tag-based web page analysis. This method extracts content based on the unique tags of each part of the web page. This method can achieve the effect of visual extraction. The extraction of the content on the web page is very accurate. It can provide users with higher retrieval accuracy and help users find the most needed web page information. Web search collects information on a specified website in real time and saves it incrementally for local use. Users can either search for web pages by title or search web pages based on content. For the obtained search results, the user may choose to sort the search results according to the relevance or timeliness of the webpages. At the same time, when the search results are displayed, the user is provided with the web page size, the time of publication, and the source of the information. Because web pages often republish, the system provides the same information merge functionality.

Web search consists of four modules: searcher, analyzer, indexer, and retriever. The four modules work together under the coordination of a main program and provide users with a stable and accurate web search engine.

Web search workflow:

1) Take out the seeds from the seed site news_site table and download the seed pages;

2) Extract the web page links from these downloaded web pages and insert them into the news_url table;

3) Select all web links from the news_url table and download them all locally. Then extract the title, time, and content according to the tags and write them into the .txt text;

4) Index the extracted .txt text content;

5) Provide search results by content and title while sorting by relevance or timeliness.

5.1 Web Searcher

The web searcher workflow is as follows:

1) Read the local site url.txt file and insert the URL of its saved torrent site into the news_site table;

2) Select all the seed URLs from the news_site table and grab the web page content corresponding to the URL. At the time of crawling, crawling was performed using two connections. The first connection is to obtain the URL session information, and then the second connection is to use this session information to establish a connection. In this way, it is possible to avoid different session information each time the same web page is accessed, and the session information of the link URL extracted from it is also different, so that it is judged as a different URL and stored repeatedly;

3) For the character stream read from the network, according to the monitoring result of the encoding detection class, the corresponding encoding format is used for decoding;

4) Insert the URL of the web page into the news_url table according to the field value in the database table that needs to fetch the web page list. According to its sub-branch field value to find the website under the theme of the website link, insert into the news_site table;

5) Select all pages (url_update=0) that have not been crawled from the news_site table, crawl them consecutively, and save them after decoding;

6) Update the contents of the database corresponding to the URL and fill in the database table with the size and status of the web page;

7) Check whether there is this URL in the outer_url table of the external network. If it does not exist, insert the URL into the outer_url table and set the corresponding field accordingly. Then save the html page in the corresponding folder of the external network.

5.2 Web Analyzer

The web analyzer workflow is as follows:

1) Acquire the content of the seed webpage according to the webpage provided by the searcher and obtain the content such as the date and title of the general webpage;

2) When extracting the seed page, certain restrictions are required and invalid web links are filtered. For filtered links, a conversion is required. Because there are many links that are relative links, they need to be converted to absolute addresses that crawlers can crawl. At the same time, when extracting web page links, it is also necessary to judge whether it is necessary to extract the corresponding release time at this time. If it is, it is extracted, and if it is not extracted, it is put into the current time;

3) When extracting an article on a web page, it is compared based on the link word extracted when the system extracts the web page tag, and the most suitable word is selected as the web page title. For date extraction, if the date can be extracted, the extracted date is put into the database, and if the date is not extracted, the current date is put into the database. For the extraction of the main body content of the webpage, we examine whether the content is included, if it is in direct addressing, content is indirectly obtained; if it is normal addressing, normal extraction is performed, and after the content is obtained, a label is filtered;

4) After extracting the title, date, and body content of the web page, write them into the local url_id.txt file, which is stored under the folder named first_crawler_time.

5.3 Web Indexer

The web indexer workflow is as follows:

1) For small webpages with docsize=0 or webpage content <30, directly set to already indexed;

2) Update the index url_update corresponding to the normal webpage to create an index;

3) Open the index file and perform a query based on the title of the web page being indexed. The condition of the query is that the word appearing in the title and the content must be included. The proportion of words that must be included is not less than 0.75, and no less than 4 words;

4) Compare the titles of the retrieved web pages. If there is less than 50% of the same content, it will not be processed. Otherwise, merge these into the same web page and reset their count values. Ordinary webpage count=0, and the same webpage has the same count value.

5) Indexing is based on the information obtained from the database and obtained count information indicating the same number of pages. It mainly establishes index fields such as url_id, url, date, crawler_date, path, title, contents, length, and source.

5.4 Web Retriever

Web search provides user title search and full-text search. Sort by relevance, and sort by time. The web retriever workflow is as follows:

1) When the user enters a search term, the user is provided with a smart prompt according to the content of the system's existing popular dictionary. Segmentation is performed when the search terms entered by the user require word segmentation services. At the same time, the user's search term is updated to the popular dictionary;

2) Search in accordance with the principle that the words entered by the user are important words. The search is based on the user's choice of using the search from the title, or from the content (default), or according to the relevance, or according to time;

3) The search function combines the search conditions based on the search requirements entered by the user and then performs the search. The search result is filled in the result class, and at the

same time, an abstract is generated based on the user's search term and the content of the retrieval web page. The abstract is then encapsulated into the result class and returned to the display page for display. The formation of the abstract is to first carry out the clauses, then segment the words after the clauses, and sort the sentence according to the number of the search keywords contained in each sentence after the segmentation. Select the sentence containing the most keywords to form the abstract. The number of abstracts is limited to 160 words;

4) In the interface, the user is provided with a title, an abstract, a size, a release time, a source, and the same number of pieces and the same display entry. Vary sorting methods to sort the search results. Users can also choose other search methods to search. If it is a web page title search, the user is not provided with abstracts, but each page provides the user with 20 search records. The index of the web page is shown in Table 1 below:

Table 1: The index of the web page

Domain name	Field.Store	index
Urlid	YES	UN_TOKENIZED
url	YES	UN_TOKENIZED
Date	YES	UN_TOKENIZED
crawler_date	YES	NO
Path	YES	UN_TOKENIZED
Title	YES	TOKENIZED
Contents	YES	TOKENIZED
Length	YES	NO
Source	YES	NO
Count	YES	UN_TOKENIZED

6. APPLICATION AND PERFORMANCE

SEfBIDm was deployed, tested, and operated by 69 branch decision-making agencies at the world’s 10th largest bank. Decision-makers from these decision-making bodies believe that SEfBIDm is rooted in the banking industry and that it is supported by banking industry knowledge and experts, which are not available in a GSEs. The information obtained from SEfBIDm has three distinct characteristics: the support of comprehensive information before decision making; the timeliness of feedback information during the execution of decisions; and the very accurate evaluation information obtained when the decision is executed.

6.1 Comprehensive Information

SEfBIDm focuses on professional websites and professional databases of the banking industry, combines the banking industry knowledge we have accumulated over a long period of time, and strives to provide decision-makers with comprehensive banking industry information and development trends through the Dashboard visualization. It ensures that decision-makers can receive comprehensive information before making decisions. This will enable decision-makers to make more objective, sustainable, and enforceable correct decisions with the help of this comprehensive information.

6.2 Timeliness Information

The vast majority of information obtained by SEfBIDm is structured information and metadata. Structured information and metadata means that the amount of information is small and the data is more concise. Therefore, it can ensure that the Decision-makers receive timely feedback on the implementation of the decision. This will allow decision-makers to confidently implement established decisions or make necessary corrections to their decisions based on implementation. SEfBIDm can provide decision-makers with daily briefings, weekly newsletters, monthly newsletters, quarterly briefings, and annual briefings on the overall situation of major banks and banking industries in the world. At the same time, special 24-hour briefings can also be provided for special events or emergencies in the banking industry.

6.3 Accuracy Information

SEfBIDm can collect, clean, and analyze the vertical chain involved in the decision-making activities and the horizontal aspects involved in the decision after the implementation of the decision. Then it can obtain accurate, professional and detailed assessment reports under the support of banking industry experts. This report allows decision-makers to objectively recognize and evaluate the decision-making effects of their decision-making behavior, thus completing the closed-loop Plan-Do-Check-Act (PDCA) of the entire decision. At the same time, it can also provide a reference for the next decision.

7. ACKNOWLEDGMENTS

We acknowledge financial support from the Chinese Scholarship Council (CSC).

8. REFERENCES

[1] Qinghua Zhu, Jia Tina Du, Fei Meng, Kewen Wu, Xiaoling Sun,(2011), Using a Delphi method and the analytic hierarchy process to evaluate Chinese search engines: A case study on Chinese search engines, Online Information Review, Vol. 35 Issue: 6, pp.942-956.

[2] MO Qian, ZHANG Shu, WANG Fang, (2012), Design and implementation of domain-oriented intelligent search engine. Computer Engineering and Applications, 48(21):112-117.

[3] Gail Steptoe-Warren, Douglas Howat, Ian Hume,(2011), Strategic thinking and decision making: literature review, Journal of Strategy and Management, Vol. 4 Issue: 3, pp.238-250.

[4] Stuti Saxena, Tariq Ali Said Mansour Al-Tamimi,(2017), Big data and Internet of Things (IoT) technologies in Omani banks: a case study, foresight, Vol. 19 Issue: 4, pp.409-420.

[5] Nadjla Hariri, (2013), Do natural language search engines really understand what users want?: A comparative study on three natural language search engines and Google, Online Information Review, Vol. 37 Issue: 2, pp.287-303.

[6] Chantola Kit, Toshiyuki Amagasa, Hiroyuki Kitagawa, (2009), Algorithms for structure-based grouping in XML-OLAP, International Journal of Web Information Systems, Vol. 5 Issue: 2, pp.122-150.