Expert Systems with Applications 37 (2010) 6515–6530

Contents lists available at ScienceDirect

Expert Systems with Applications

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

An Intelligent information segmentation approach to extract financial data for business valuation

Jia-Lang Seng[a,*,1], J.T. Lai[b]

[a] Dept. and Graduate School of Accounting, College of Commerce, National Chengchi University, Taipei, Taiwan
[b] MICRONIX Technology Inc., IT Division, Taipei, Taiwan

ARTICLE INFO

Keywords:
Intelligent information extraction
Information retrieval
Word segmentation
Knowledge base
Text processing
Natural language processing
Business valuation
Financial data
Non-financial data

ABSTRACT

Due to an increase in the wealth of electronic resources on the Internet in the past several years, the birth of the search engine has brought the utmost convenience and efficiency for users. However, searching for data by keyword retrieval techniques in information retrieval is not contented with some users' specific needs due to a large number of network resources and users on the Internet. Information extraction is an improvement method which extracts the important specific event or produces specific relations among information from documents. Information extraction can not only filter unnecessary information in any documents but also produce specific important messages and summaries that users are interested in.

Business valuation is collecting, analysis, and applying to financial or non-financial integral information to appraise the business value. The evaluated results are used in the commerce pricing for the business decision and intangible assets. There are specific information and events about business valuation stored in the Intelligent financial statements, notes to financial statements, and financial news of Taiwan's companies at present and data is presented by the HTML and PDF files. Hence, we developed an information extraction system of Chinese financial data for business valuation from the domestic business financial statements, notes to financial statements, and financial news as the data sources. We extracted the correct financial data and their corresponding Business Valuation Model to achieve an automatic extraction in the financial data from these different heterogeneous data sources. Users can collect the relevant valid valuation information and learn valuation models concepts within a very short time to improve accuracy and efficiency in text processing quality.

1. Introduction

1.1. Research motivation

Due to an increase in the wealth of electronic resources on the Internet over the last several years, the birth of search engines such as Google and Yahoo! has brought about the utmost convenience and efficiency for users. Search engines can filter unnecessary documents in the Internet resources and find proper information to meet users' needs as much as possible. However, searching for data by keyword retrieval techniques is not contented with some users' specific needs due to a large number of Internet resources on the Internet. We find that these documents retrieved by search engines usually have some irrelevant data and contents. Hence, users must research on whether the content of each file meet their specific demand and a great deal of time in the manual processing of data.

Because keyword retrieval techniques cannot meet the specific need for users, relevant new research studies have appeared constantly in recent years aiming to improve the efficiency and accuracy of information retrieval (IR). Information extraction (IE) is an improvement method to extract important specific events or produce specific relations among information from documents. Information extraction can not only filter out unnecessary information in documents but also produce specific important messages and summaries that users are interested in. Li, Wong, and Yuan (2003) noted that information extraction is a use of extension for analysis input documents by way of natural language processing techniques and capturing and storing specific information in the artificial intelligence techniques. Hence, nowadays, information extraction technique plays an indispensable role in information retrieval technique.

Business valuation is collecting, analysis, and applying to financial or non-financial integral information to appraise the business value. The evaluated results are used in the commerce pricing for business decisions and intangible assets. There is specific information and events for business valuation stored in the Chinese financial statements, notes to financial statements, and financial news at present in Chinese businesses and these data sources are stored in the HTML web pages and PDF files. Nevertheless, valuation experts must collect the relevant raw data on-

[*] Corresponding author.
E-mail addresses: seng@nccu.edu.tw (J.-L. Seng), jljan2@gmail.com (J.T. Lai).
[1] Distinguished Professor & Department Chair.

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.02.134

line by themselves and then read and filter the necessary information in detail by the manual processing tasks. Therefore, they must often spend a large amount of time collecting the raw data from different data sources and there are no specific web sites or databases to make inquiries and to analyze these data about business valuation for users or experts. Hence, we study information extraction methods from these data sources financial statements, notes to financial statements, and financial news, and extract useful financial data for business valuation from these different web heterogeneous data sources. Users can collect correct financial data for business valuation within the shortest time to improve the exactness and efficiency in text processing quality. These data applied to data mining and statistic analysis will be beneficial. Valuation results are used as key indicators for carrying out business activities such as stock listings and merge and acquisition decisions in the future.

1.2. Research issue

We focus on how to use Intelligent information extraction techniques in collecting financial data for business valuation from financial statements, notes to financial statements and financial news. In the past, relevant research on financial information of some specific events such as takeovers, bankruptcy, joint ventures has been performed in the area of information extraction but there is no advanced analysis and study on the issue of business valuation. Therefore, we use the two major techniques of Intelligent word segmentation and information extraction to search for related Chinese financial data or financial events for business valuation. We pre-define keywords extracted in business valuation. The system can seek, match, and extract data we need according to definite methods. It generates exactly the highly relevant information for business valuation. After collecting the useful data, we can store the structured information into the database and develop the optimization models by data mining and statistical analysis. Users not only collect fast the correct financial data, but also show the data exactness and performance. Hence, the purpose of this research can achieve the automatic extraction of financial data for business valuation from different web heterogeneous data sources. However, research scope is only limited to the financial data. We proceed to extract these data by definite methods and collect integral data to be stored in the documents. Finally, we evaluate the integrality and exactness from these data the system extracts.

This research mainly involves the development of a Intelligent information extraction system in collecting financial data for business valuation. The data sources are financial statements and notes to financial statements from annual reports on the Market Observation Post System, and financial news on the China Times web site. The scope of the research is as follows. Information retrieval is a method of finding information we want to process by using keyword techniques common in search engines, and information extraction is an improvement method by finding specific events and information from the retrieved documents. As part of business valuation, we try to categorize these contents in annual reports. In the same way, this method can be applied to financial news. The classification has two dimensions. One dimension is financial data versus non-financial data. The other dimension is financial report versus non-financial report. This will result in four categories of data source. Our research falls in the class of financial data and financial report which represents more than half of the data population.

In order to extract financial data for business valuation from financial statements, notes to financial statements, and financial news, we address two issues in this research.

- Keyword retrieval and extraction.
  - o How to define the part-of-speech tags of the Chinese keywords we need to extract financial data from financial news and notes to financial statements. These selected keywords are segmented and analyzed first in the Intelligent keyword extraction, we design an extraction method according to pre-defined part-of-speech tags and the data attributes.
  - o How to develop a regular expression method in order to identify individual financial data record and fields in financial statements.
- Development of valuation knowledge base.
  - o We describe Business Valuation Models and illustrate hierarchy relationships among models to design the concept trees.
  - o Finally, according to the feature of the knowledge base design, we turn the concept trees into the records and fields accessible in the knowledge base.

2. Literature review

2.1. Intelligent information extraction

Information extraction is a developmental concept and a technology used to improve the performance in information retrieval. It mainly performs syntax analysis, the extraction of important nouns or phrases, and semantic analysis to data. The useful information is extracted from the miscellaneous data, allowing users to quickly understand the essential information and concept from documents. Pure-text pages on the Internet are usually viewed as non-structured documents. According to pre-defined templates, specific information is extracted from these documents. The system generates field or theme summaries which the users are interested in, or the suitable and useful domain data found from a large number of databases are extracted into structured information. Simply stated, it is how to turn it into information from the huge and various data such as noun phrases analysis, syntax analysis, semantic analysis is extracted as the important fact, helping users to understand the meaning of the information from the documents. Hence, the information extraction is a very important role in the text processing.

The information extraction techniques offer information that users are interested in. Its technology, depending on the natural language processing, can extract the specific lexicons formed as the concept or information. Natural language processing (NLP), enabling the computer to deal with and to understand people's languages used, enables the computer to understand human knowledge and enables it to make the efficient communication between people and the computer. It is usually applied to information retrieval (Baeza-Yates & Ribeiro-Neto, 1999; Chien & Pu, 1996), information extraction, the question and answer system, the document classification, the machine translation, the writing aid, the voice recognition. Intelligent word segmentation and lexicon analysis are the most important and main methods in natural language processing.

For non-structured documents, the above-mentioned Intelligent information extraction system can include three components as follows: the word segmentation module, the lexical analysis module, and the lexical extraction module.

- *Word segmentation module:* Lexicons are not separated by spaces like English in Chinese documents. In order to recognize a single lexicon effectively, the system must be designed through algorithms and programs that can separate lexicons clearly in texts.
- *Lexical analysis module:* This module is defined as generating parts-of-speech (POS) tags of these lexicons generated by the word segmentation module.

- *Lexical extraction module:* Lexicons of the word segmentation have various parts-of-speech (POS). In order to effectively extract the important lexicon, the system must be developed by some proposed models and methods to extract keywords according to the morphological feature, related lexicons, and the lexicon appearance frequency in texts.

Structured data, which are also called tabular data, simplify and standardize information in documents. Tables are the formalized presentation of miscellaneous information and with readable and clear properties. Because the fields in records represent different attributes, the user can quickly understand the meanings. For extraction of tabular data, we only understand the structure of tables and field attributes and then extract the essential information we process.

2.2. Intelligent word segmentation

Word segmentation is the most basic and also the most important processing method in natural language processing. In Intelligent word segmentation, a lexicon is viewed as a recognizable unit, not a character. However, a computer is unable to distinguish lexicons. In other words, segmentation methods must identify them correctly and fully in order to deal with the feasibility of all consequent processes.

Chinese lexicons are not separated by spaces as in English. A sentence made up of vocabularies that users input in the retrieval system must be segmented to deal with and to distinguish every lexicon. The system can make further treatments in the language process system after distinguishing lexicons, for example, machine translation, language analysis, knowledge extraction. In Intelligent word segmentation, the documents or texts are word matching with the corpus-based dictionary and important keywords are found out, but there still exist problems in Intelligent word segmentation. More specifically, word segmentation still has some ambiguous problems, for instance, the same Intelligent word sequences having different segmentation results. In order to solve this problem, high quality methods in word segmentation must be selected. In addition, unknown words are also an unpredictable problem in the word segmentation process. According to the Chinese Knowledge and Information Processing (CKIP) group in Academia Sinica, there are nearly 3–5% of unknown words in a general article. Hence, a standard technology and method is necessary in distinguishing unknown words and part-of-speech tags. Neither a dictionary nor technology can include and recognize all Chinese lexicons so far, so the solution now is to collect all Chinese vocabularies as often as possible. Most unknown words are still comparatively the majority with the named entities at present, because named entities are only limited in some files of special fields. Besides more common keywords in some specific dictionaries, a few common key words must be distinguished by technical and specific methods. Presently, the benchmark method of word segmentation is usually recall and precision (Cercone, Huang, Peng, & Schurmans, 2003).

Furthermore, in the past few decades of research on word segmentation, problems are classified into word sequences matching, unknown words, named entities and part-of-speech tagging as we will see below. We will illustrate the precise solutions to these problems in the following section.

- *Technological problems of Intelligent word segmentation*: In retrieval methods, the first step is to set up an index of words collected. There have been three approaches in traditional technology. They are the dictionary based method, the character based method, and the statistic based method (Cercone et al., 2003; Chen, Gey, He, Meggs, & Xu, 1997). The dictionary based method is the most general method in Chinese retrieval techniques in early periods and matches words quickly by the inverted method. Then, a large number of lexicons are collected and word segmentation is based on heuristic rules (Chen et al., 1997; Chen & Ma, 2005). The common heuristic rules are, for example, the maximum matching, the word length, the morphemic, and the probability four methods. In these four methods, the maximum matching is usually used for solving the ambiguous problems.
- *Words length:* We segment the input sentence into partitions and finds the minimum standard deviation after every divided length. The formula of the minimum standard deviation is $(L(W1) - \text{Mean})2 + (L(W2) - \text{Mean})2 + \cdots + (L(Wn) - \text{Mean})2$. $W1$ to Wn are partial word terms after the sentence is segmented and obtains their length.
- *Morpheme*: The sentence is segmented by morphemes. The morpheme refers to the minimum unit of the meaning or the grammar.
- *Probability*: The probability is similar to a method of statistical models, but the difference is that probability rule is solving the ambiguous problems after word segmentation. The statistical models can be used in whole inputting sentences and deciding the calculation of the optimization to find the appropriate segmentation result.

2.3. Part-of-speech tagging model

The function of part-of-speech (POS) tagging is mainly applied to eliminate the segmentation ambiguities and offering the correct meaning. The disambiguation method is used for the contextual meaning to recognize the goal meaning. In a Chinese word, the same lexicon has quite different meanings, but it usually shows a specific meaning on the certain sentence structure. In order to improve the correct rate of information retrieval, part-of-speech tagging is proofread with the manual method in the early days, but this is unable to improve both efficiencies and qualities. It must spend a large amount of manual processing tasks and time, and it also is unable to produce a large number of texts. With the progress of information science and technology, such a concept is applied to the information system and utilizes the statistical method showing parts-of-speech tagging of each lexicon. Under such a precondition, as it is necessary to raise the accuracy rate of word segmentation, the quality of part-of-speech tagging can be improved.

2.4. Keyword extraction on text data

Keyword extraction is essential because it represents certain concepts in text documents. Due to all words having different levels of importance in text documents, these research issues concerning keyword extraction have been discussed constantly in recent years. However, the above-mentioned word segmentation is a method of extracting lexicons via modern technologies. For the user, it is only a method of extracting lexicons via recognition modules, and this research views the keyword extraction as a concept-based or a message-based extraction. The concept-based extraction can extract the most important keywords to present significant concepts and enable users to understand the important information in text documents. Estimating how to choose the meaningful keyword is essential. In general, what is called a meaningful keyword is composed of nouns and verbs, and stored in documents. Because of indicating independent facts, most nouns can be accurate descriptions in concepts. But as for other words, for instance, adjectives, prepositions, and adverb modifiable words, they are not very important. Hence, most modifiable words can be filtered out through keyword extraction in advance. Besides, keywords are covered under the situation with quite high frequency of repeated appearances and can be viewed as the important concepts in the document. Therefore, we can calculate the frequency of all words, and the keywords selected with high frequency represent important concepts.

The co-occurrence analysis is two or more same lexicons appearing together in the same document at the same time. In order to reduce the search time, the co-occurrence keywords appear as different lexicons limited to the same sentence and are able to extract the lexicon quickly. The high-related lexicons with the specific field in a common document are called the field association lexicons. For instance, we study Account Names with features of field association lexicons in financial statements, notes to financial statements, and financial news. There has been a method in the field association lexicons (Atlam, Fuketa, Kashiji, Nakata, & Aoe, 2002). It is applied to use lexicons extracted to perform the document classification. This method is that the co-occurrence keywords are offered the weight value, and then these keywords with the number of many classifications are reduced, so as to ensure document classification with high efficiency and exactness.

The dictionary based analysis has the highest accuracy of extracting lexicons in three methods. The method searches and extracts lexicons via word matching from the dictionary base. This can ensure that the lexicons are fully correct. But the shortcoming is limited to the scale from the source. It is almost impossible to include all lexicons. However, with such huge Internet resources, keyword matching we only depend on cannot filter ineffective information. In recent years, creating the corpus of related lexicons has been increasing sharply. Under the semantic web (SW) concepts, the dictionary with and among using high-related lexicons are integrated. When inquiring a keyword, the system cannot only receive the data about the keyword, but also relevant concepts about the lexicon to improve the data quality. The synonymous dictionaries were set up automatically.

2.5. Structural extraction on tabular data

The forms of Internet resources are diverse. Resources of the Internet are like a large-scale fictitious database to store structural or nonstructural data in various fields. The technological development of information retrieval and extraction is not limited to specific users due to extension of the need. The retrieval use of Internet resources relies on search engines at present. Search engines usually use web pages as the search target. After users have logged in to the web sites, the search engines can process and convey the specific web pages by keywords searching. The search technology can construct a web map in the system at present and then use robots or spiders to extract the web pages on a periodical schedule. The web page will be stored in the database.

However, spiders only help users look for associated documents from network resources. They cannot make further extractions to express certain important concepts from contents in documents. For structured information extraction techniques on web page, several research studies have been proposed constantly in recent years. Zhai and Liu (2005) proposed a method consisting of two steps from web information extraction. The first step uses an improvement to the previous Mining Data Records (MDR) algorithm, identifying individual data records from a web page that contain structured data records. The second step uses a novel partial alignment technique to match corresponding data items and fields from data records. The experiment results of two steps show there are more accurate than the previous methods. Limanto, Giang, and Trung provided the web page-based information extraction engine. The main structure components of an information extraction engine include wrapper generator, wrapper database, and extractor component.

Wrapper generator component (Limanto et al., 2005; Lochovsky & Wang, 2003) provides the content analysis of web pages which found by spiders. The content analysis discovers the repeated patterns in a page and can generate the regular expression. For example, the tabulation data in a certain web page, repeatedly covers these tags <tr></tr> or <td></td>. These HTML tags can be put into the regular expression. Lochovsky and Wang (2003) noted a C-repeated pattern method that induced a regular expression from the repeated substring. This method is viewed as HTML tags as one token and creates a token suffix tree to find out C-repeated patterns. During the process of tracing the token suffix tree, it is viewed as a part of the C-repeated patterns to obtain the difference between child nodes in same father node to equal the length of the pattern. This pattern is regarded as C-repeated patterns, then, the partial patterns are kept back first, and then rebuild a token suffix tree from left HTML tags to find out new patterns recursively until no new patterns appear. The final result of the regular expression in the HTML-page will be stored in the wrapper database. The system can extract the correct regular expression fit with the web page from the wrapper database and acquire the correct information from the web page similar to a certain format.

In addition to the above-mentioned HTML tags files, PDF is a commonly used and presented Internet resource. About structural extraction on PDF documents, relevant research studies have reported constantly in recent years. Liu, Mitra, Giles, and Bai (2006) proposed how to extract tabular data and to convert PDF files into TXT files. The method is composed of three steps: converting a PDF document into a formatted text file, detecting the table candidates based on location analysis and keyword matching, and recognizing the table structure. Rosenfeld, Feldman, and Aumann (2002) proposed a general procedure for structural extraction, which allows for automatic extraction of data from a document based on their visual characteristics and relative position in the document layout. In other words, we use an automatic process module that accepts a formatted document as input, and returns a set of pre-defined attributes or elements of the document, each assigned to a corresponding field, e.g. "AUTHOR = . . ., TITLE = . . .,". The set of field names and document elements that get assigned to them is problem-dependent, and may be different for different types of documents. Thus, we search a system that learns how to extract the proper document elements based on examples provided by a domain expert.

In summary, this section provides important literature review and theoretical basis of information extraction techniques and Business Valuation Models. Information extraction is the use of extension in information retrieval techniques using natural language processing techniques such as the methods of Intelligent word segmentation and part-of-speech tagging. These methods are applied to the area of information extraction to improve the exactness and efficiency in text processing quality. In Business Valuation Models, we discuss three common methods first and introduce the use of parameters in each model so as to understand corresponding relationships between the financial data and valuation models later. Then, we try to take advantage of the research issues we mention which is applied to extraction methods for financial data of business valuation and a design of valuation model knowledge base and to develop a set of concrete research models from heterogeneous data sources.

3. Research approach

This section presents mainly how to develop the research model for Intelligent information extraction in business valuation. We use the financial statements, notes to financial statements, and financial news as the data sources and search for financial data for business valuation. The financial data are composed of organization names, specific times, Account Names and money or percent. Users and financial experts can collect and analyze correct financial data from these sources within the shortest time. We illustrate the research model in Fig. 1 as follows. The figure illustrates each main process in the research model to achieve the research objective. We develop different extraction methods according to different data sources.

3.1. Information extraction on financial statements

Financial statements are formal records of a business's financial activities. Financial statements can be classified into balance sheets, income statements, cash flow statements, and owner's equity statements. We illustrate features of these financial statements and the corresponding relationships between Account Names and Business Valuation Models in the following section.

The financial statements stored in web pages are shown by HTML documents. Beside organization names and time we extract in the table head, the tabulation structure is usually shown in the financial statements. In HTML tags, table tags are composed of such necessary tags as "<table></table>", "<tr></tr>", "<td></td>". We must check and patch incomplete tags. These HTML tags are an important element of the information extraction (Krupl, Herzog, & Gatterbauer, 2005).

HTML tags must be distinguished to be shown correctly through the Internet browser. This is a file with tabulation tags and data. HTML tags in pairs can define the specific range usage and enable the data to achieve their effect in the use of specific ranges. For instance, the necessary tags set up of the table are "<table></table>". "<table>" is the start tag; "</table>" is the end tag. These two tags are the most basic important elements while the table is set up. All other table tags must be placed between the two tags and can perform their performance. The tags "<tr>···</tr> are defined as the row in order to show a record. In other words, the number of records in the table depends on the number of "<tr>···</tr>" placed within "<table>···</table>". The paired tags "<td>···</td>" are defined as a field in a record. The tags "<th>···</th>" are defined as the title name in the field. These tags "<td>···</td>" and "<th>···</th>" must be placed within "<tr>···</tr>". After we write these tags in an HTML document, the document is distinguished through a web browser.

Generally speaking, the features of records can be understood clearly by the tabulation data set up. We receive important information rapidly in the records. In financial statements, the tabulation data are shown clearly to express financial condition in a business's financial activities. The users can collect and understand the information or messages about financial activities. In order to enable the system to extract data accurately, for example with the specific web pages, we illustrate the method of information extraction.

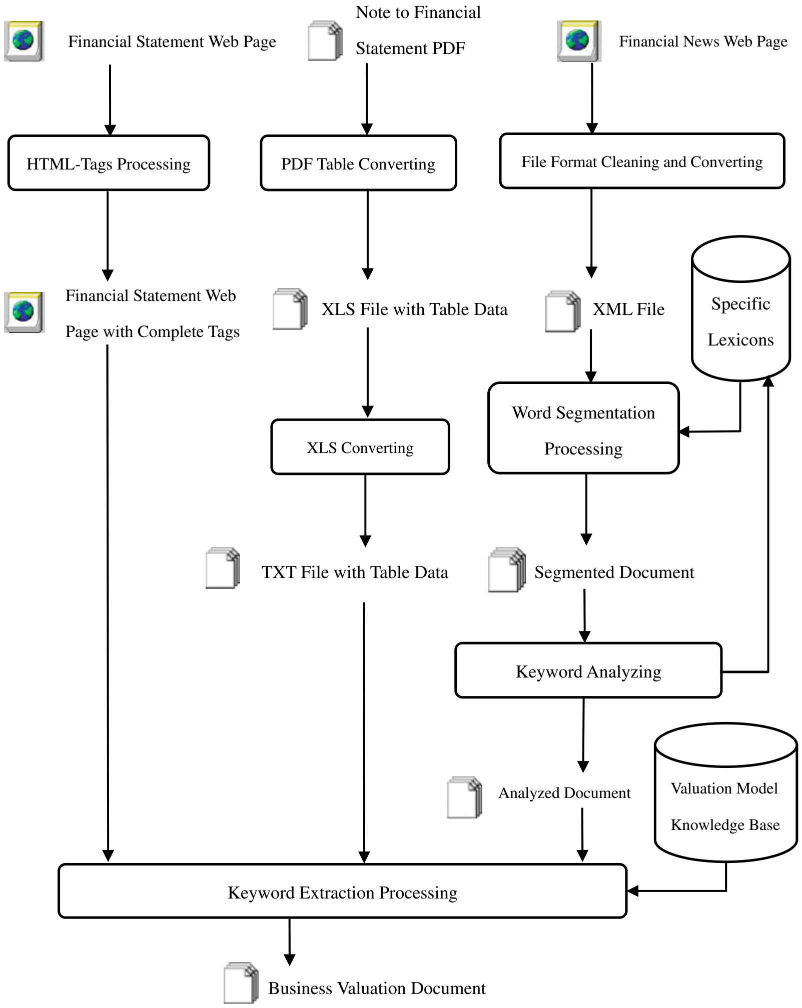

Financial Statement Web Page
Note to Financial Statement PDF
Financial News Web Page
HTML-Tags Processing
PDF Table Converting
File Format Cleaning and Converting
Financial Statement Web Page with Complete Tags
XLS File with Table Data
XML File
Specific Lexicons
XLS Converting
Word Segmentation Processing
TXT File with Table Data
Segmented Document
Keyword Analyzing
Analyzed Document
Valuation Model Knowledge Base
Keyword Extraction Processing
Business Valuation Document

Fig. 1. Research structure.

In a HTML web page, no matter whether it is HTML tags or original data, it can be regarded as these data objects or the string instances. Regular expression (RE) is an expression method of word sequences. RE can briefly express word sequences with specific or miscellaneous characteristics and find out its regular characteristic from complicated word sequences. We attempt to look for the rule of the syntax in HTML tags as word sequences. When searching for the Account Name viewed as the parameter of word sequences, we use the parameter to search for model names and parameters in the database.

3.2. Information extraction on notes to financial statements

Financial statements describe the numeric data records. Notes to financial statements disclose events that the financial statements cannot illustrate and describe by quantitative or descriptive methods. Notes to financial statements are composed of events, for example, accounting policies, contents of significant accounts, related party transactions, and disclosure events. We use important contents of significant accounts as the data sources to extract financial data for business valuation.

In the notes to financial statements, contents of significant accounts are shown in tables. For example, cash in Account Name is composed of cash on hand, saving deposits, fixed deposits, presented in the table structure. Notes to financial statements are shown in quarterly reports and annual reports that are stored in PDF files. Hence, extracting table data from PDF files is the research issue. In this research, we use the specific PDF converting tool to convert tables in PDF files to TXT files.

We use notes to financial statements as the data source and convert a PDF file into a TXT file to extract named entities including Account Names, time and money or percent. In the extraction method, the system module extracts all named entities by word morphemes in a TXT file to recognize POS tags for each field. Regarding Account Name, we search for its corresponding model names in the valuation model knowledge base.

3.3. Information extraction on financial news

The financial news we selected must include the information with business valuation, for example, issued financial reports, stock price information, and operating activities about listed companies every year.

Financial news is one of the data sources and stored in HTML or XHTML formats. In order to meet the need and the specification in the Intelligent word segmentation tool, the first step is to eliminate extra information such as unnecessary labels and notes and the specific file stored in the file. Chen and Ma (2001) provided an operation method about the corpus construction and the word segmentation tool in Fig. 2. After loading the XML file, the content in the file is segmented by the built-in dictionary based method. This specific lexicon base in Fig. 3 is classified into built-in and self-constructed two functions. But the ambiguous problems are still generated. What is known as the ambiguity has many different segmentation groups.

3.3.1. Account name analysis

In order to extract the Account Names exactly, we use the self-constructed dictionary function in the segmentation tool. The self-constructed dictionary function is the most simple function and achieves the string matching exactly. The number of the Account Names has a certain quantity. Tt is not essential to waste the manual power and time while setting up the dictionary.

When Account Names are set up in the specific dictionary, we must consider some synonyms. In the general Chinese financial news, abbreviations and synonyms derived from Account Names are widely used. Although the lexicons are always colloquial language that most people fully understand, the segmentation system is open to confusion and recognition mistakes. Hence, the synonyms about the Account Names are included while the dictionary is set up. Finally we define “Acc” as a new POS tag of Account Names. The new tagging is viewed as a basis of information extraction.

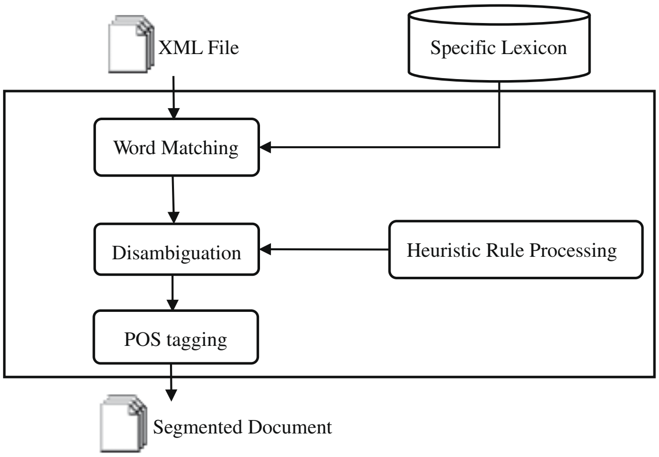


Fig. 2. Word segmentation model.

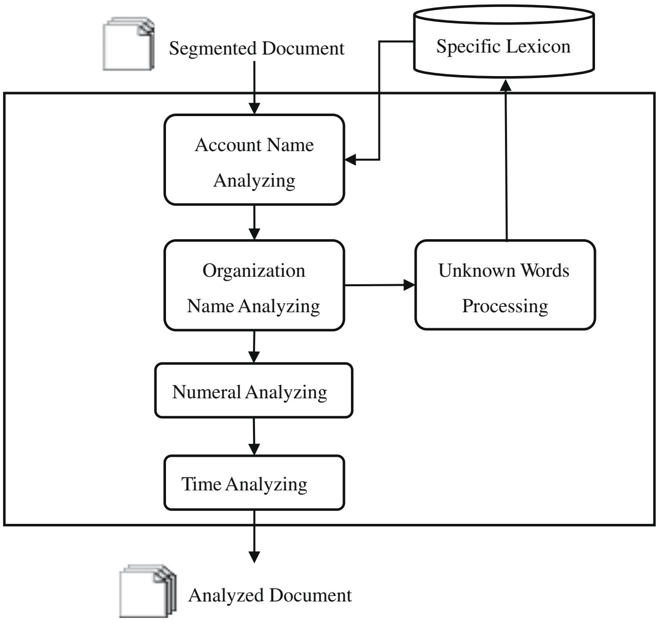


Fig. 3. Word analysis model.

3.3.2. Organization name analysis

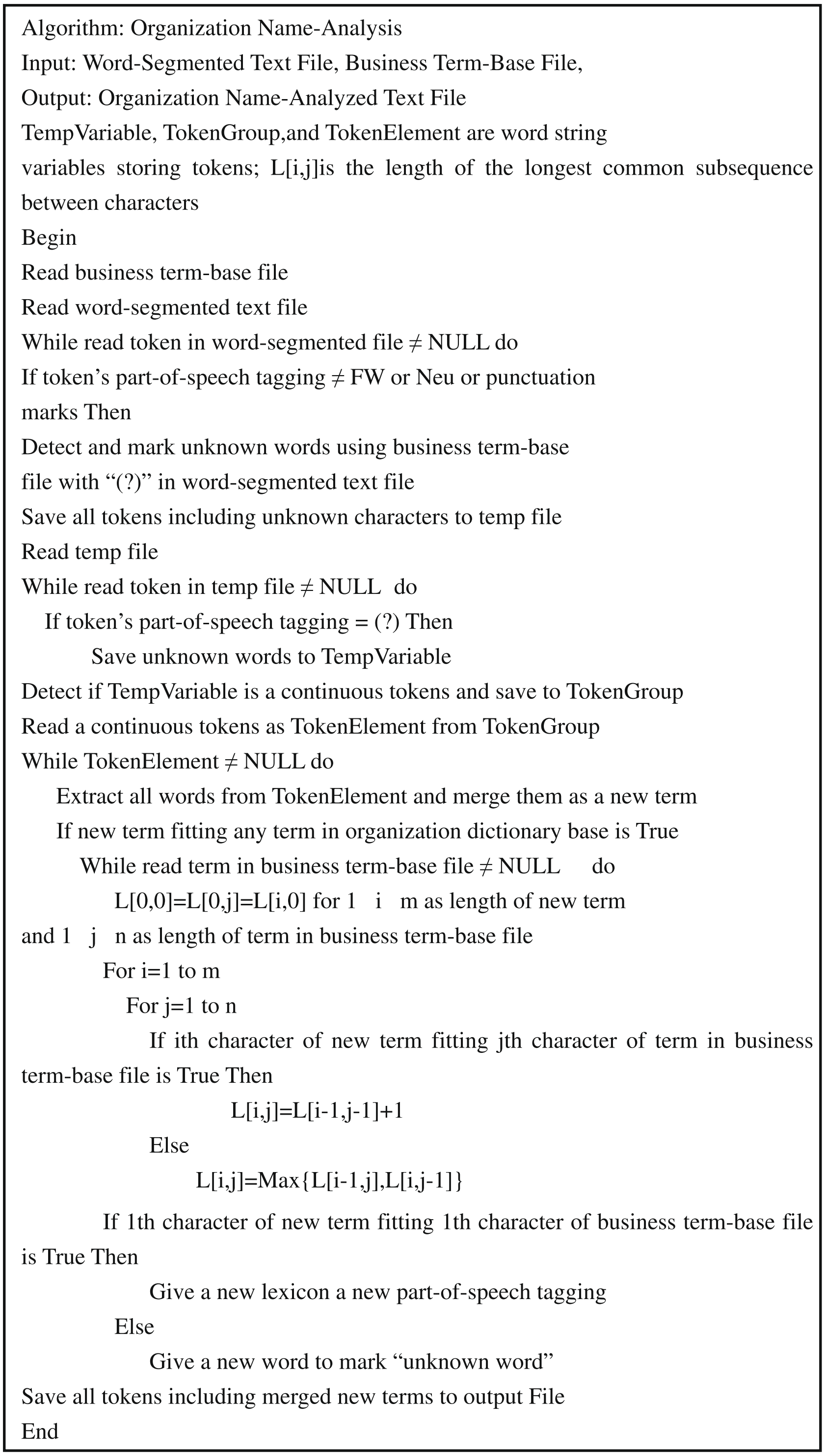
In organization names, the title of financial statements must be requested to write the full name of an organization or a company, yet in general financial news or non-financial statements, organization names are showed by their abbreviations. Organization names are defined originally as named entities, but under different organization names, there are many inappropriate noun classifications. This situation is open to causing low quality and accuracy.

In order to solve the above-mentioned problems, Fig. 4 shows an organization name analysis algorithm. First, formal organization names will be collected at the specific dictionary. According to the named entity dictionary, the analysis algorithm detects first an input text if keywords with organization names are existent. If it is, we use the symbol "(?)" to express the candidates selected by the specific dictionary. All input text contents or documents are financial news and these organization names are shown as the abbreviations in these financial news.

In addition, when we name these abbreviations, most names are regarded as a part of their abbreviations by selecting the first Chinese word of the full names. After finishing all character word sequences matching with each other, we must check its position for raising accuracy. Finally, when the abbreviation can meet the above-mentioned request, the system tags the new part-of-speech. The new part-of-speech is viewed as a term of information extraction and stored in the specific lexicon base.

3.3.3. Time and date analysis

Time is applied to a period or time point of financial activities during a certain fiscal year. Although the Chinese parser can solve some of the above-mentioned problems, the recognition still has



```
Algorithm: Organization Name-Analysis
Input: Word-Segmented Text File, Business Term-Base File,
Output: Organization Name-Analyzed Text File
TempVariable, TokenGroup,and TokenElement are word string
variables storing tokens; L[i,j]is the length of the longest common subsequence
between characters
Begin
Read business term-base file
Read word-segmented text file
While read token in word-segmented file ≠ NULL do
If token's part-of-speech tagging ≠ FW or Neu or punctuation
marks Then
Detect and mark unknown words using business term-base
file with "(?)" in word-segmented text file
Save all tokens including unknown characters to temp file
Read temp file
While read token in temp file ≠ NULL  do
  If token's part-of-speech tagging = (?) Then
      Save unknown words to TempVariable
Detect if TempVariable is a continuous tokens and save to TokenGroup
Read a continuous tokens as TokenElement from TokenGroup
While TokenElement ≠ NULL do
    Extract all words from TokenElement and merge them as a new term
    If new term fitting any term in organization dictionary base is True
      While read term in business term-base file ≠ NULL     do
        L[0,0]=L[0,j]=L[i,0] for 1   i   m as length of new term
and 1   j   n as length of term in business term-base file
       For i=1 to m
         For j=1 to n
           If ith character of new term fitting jth character of term in business
term-base file is True Then
                L[i,j]=L[i-1,j-1]+1
           Else
               L[i,j]=Max{L[i-1,j],L[i,j-1]}
       If 1th character of new term fitting 1th character of business term-base file
is True Then
           Give a new lexicon a new part-of-speech tagging
        Else
           Give a new word to mark "unknown word"
Save all tokens including merged new terms to output File
End
```

Fig. 4. Organization name analysis algorithm.

some obscured questions existing in the parts-of-speech. Generally speaking, the above-mentioned continuous segmented time words must be regarded as a specific time lexicon. Fig. 5 shows a time analysis algorithm. In the segmentation article, we first detect parameters with the POS tags of time morphemes. In other words, the algorithm must detect continual segmented word groups with the POS tags of time morphemes and mark their absolute position. The new POS tag is viewed as the term of information extraction.

3.3.4. Monetary and percentage analysis

In money and percent, there are some existent problems in financial news after word segmentation. Due to existing incorrect word segmentation results in money and percent, the CKIP groups in the Academia Sinica developed the Chinese segmentation tool to solve ambiguity so as to achieve the grammar rationality, but there are still problems in the POS tag. In other words, there are still incorrect POS tagging problems. Hence, we not only try to extract correctly the POS tag-based data we need, but also make continuous segmented words form a complete lexicon. Fig. 6 shows a money and percent analysis algorithm. The continuous segmented words are merged into a lexicon in the bottom line part and offered a new POS tag Financial Dollar Sign (FDS). The new POS tag is viewed as a term of information extraction.

3.4. Keyword extraction on financial news

The POS tags are to determine if the word segmentation file after keyword analysis is extracted correctly to collect the financial data for business valuation. The financial data are composed of four concepts including organization names (Corp), Account Names (Acc), time (Time), and money or percent (FDS). Except for extracting the essential four named entities we defined, in the information extraction literature the verb is selected as an important factor in the keyword extraction. The paragraph in word segmentation files is also viewed as the boundary between data. So the definition of the keyword extraction is that the financial data for business valuation are composed of the four above-mentioned named entities and verbs. The above-mentioned conditions are the thoughts on designing the algorithm. However, time lexicons are not usually shown in the news about financial data of business valuation. Even though the time lexicon appears in the financial news, it is uncertain if the lexicon is viewed as an attribute of valuation data. Hence, in the extraction process, the system can still accept the data with only other three named entities. Fig. 7 shows an information extraction algorithm on financial news.

3.5. Valuation model analysis

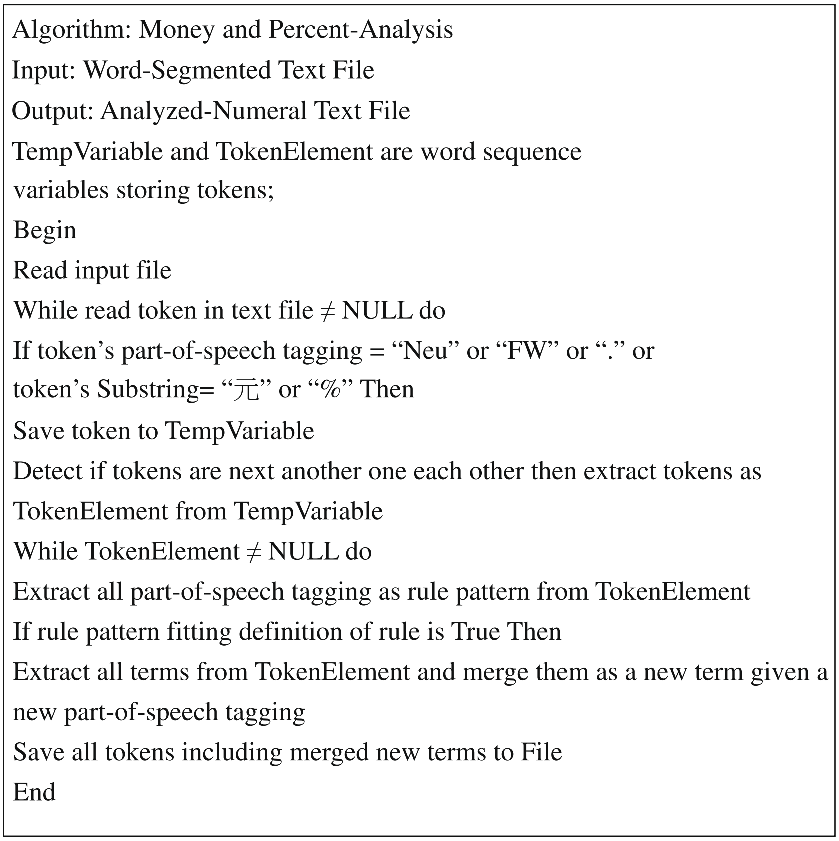
3.5.1. Business valuation

Business valuation is a money figure or interval value calculated for a valuation purpose. Because of the abstract and the highly uncertain speciality, business valuation is a highly difficult business technology. Valuation performed should confirm the valuation purpose. Simply stated, it is what is carrying out the economical behavior to the valuation purpose. Because of different valuation purposes, the types of considerable future benefits and risks may be different. Under this condition, experts should consider how to adopt applicable valuation models. Once confirming the valuation purpose, experts can correctly estimate the range, basic data, and parameters taken in about business valuation. The more the professional knowledge experts possess, the more clearly the development of the valuation theory is. Once data is more reliable, valuation work is more inclined to the scientific process. Valuation results can be quantifiable and verifiable.

In the financial theory, there are many methods and models about business valuation. We divided them into three common usage approaches: the income-based approach, the market-based approach, and the asset-based approach. The income-based approach pays attention to get the present value via the discount rate for creating future income flow or to making the business value via the economic added value. The market-based approach utilizes the value multiples of analogy companies such as earning multiples, book value multiples, and revenues multiples. The business value is inferred from the multiples. The asset-based approach is ad-

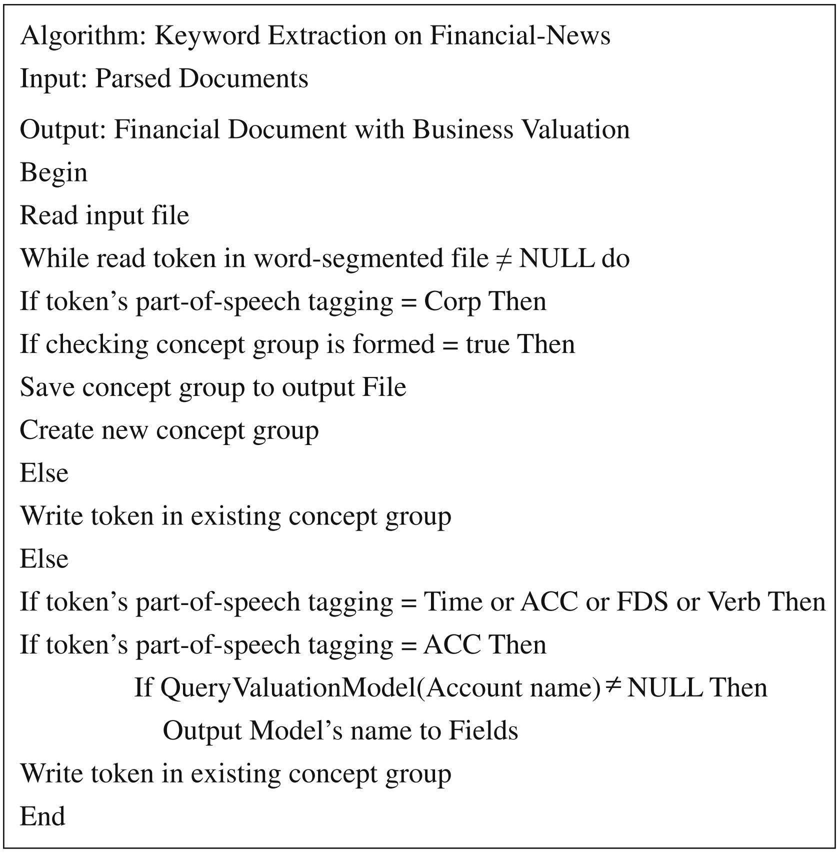
```
Algorithm: Time-Analysis
Input: Word-Segmented Text File
Output: Analyzed-Time Text File
TempVariable, and TokenElement are word sequence variables storing tokens;
Begin
Read input file
While Read token in text file ≠ NULL do
If Token with time morphemes is true = Then
Save token to TempVariable
Detect if tokens are next another one each other then extract tokens as TokenElement from TempVariable
While TokenElement ≠ NULL do
Extract all part-of-speech tagging as RulePattern from TokenElement
If Token- Element existing certain specific time morphemes is True Then
Extract all words from TokenElement and merge them as a new lexicon given a new part-of-speech tagging
Save all tokens including merged new terms to output File
End
```

Fig. 5. Time analysis algorithm.



```
Algorithm: Money and Percent-Analysis
Input: Word-Segmented Text File
Output: Analyzed-Numeral Text File
TempVariable and TokenElement are word sequence
variables storing tokens;
Begin
Read input file
While read token in text file ≠ NULL do
If token's part-of-speech tagging = "Neu" or "FW" or "." or
token's Substring= "元" or "%" Then
Save token to TempVariable
Detect if tokens are next another one each other then extract tokens as
TokenElement from TempVariable
While TokenElement ≠ NULL do
Extract all part-of-speech tagging as rule pattern from TokenElement
If rule pattern fitting definition of rule is True Then
Extract all terms from TokenElement and merge them as a new term given a
new part-of-speech tagging
Save all tokens including merged new terms to File
End
```

Fig. 6. Money and percent analysis algorithm.



```
Algorithm: Keyword Extraction on Financial-News
Input: Parsed Documents
Output: Financial Document with Business Valuation
Begin
Read input file
While read token in word-segmented file ≠ NULL do
If token's part-of-speech tagging = Corp Then
If checking concept group is formed = true Then
Save concept group to output File
Create new concept group
Else
Write token in existing concept group
Else
If token's part-of-speech tagging = Time or ACC or FDS or Verb Then
If token's part-of-speech tagging = ACC Then
            If QueryValuationModel(Account name) ≠ NULL Then
               Output Model's name to Fields
Write token in existing concept group
End
```

Fig. 7. Information extraction on financial news algorithm.

justed and re-estimated according to each account of balance sheets. The business value is inferred from the difference of the assets and debt after adjusting. Owing to consider contents of the specific data sources and practical need, under this situation we discuss the following and necessary valuation models.

3.5.1.1. Income-based approach. The discounting of the income-based approach considers the cash flow values at different time points. The most important two factors in the discount rate are predicted for the future income and the estimation for the risk adjustment in the income-based approach. The approaches estimating the present value are divided into two, the capitalization and the discounting. The discount rate is the return that changes a series of future income into the current value. The capitalization rate is the divisor that changes single specific future income into the current value. We illustrate two different future benefit variables: Discounted Cash Flow (DCF) and Economic Value Added (EVA).

Under the income-based approach, the value calculation is the flow of the future income discounted by the discount rate. DCF is the most theoretical specialty and likely to be the most correct valuation method. The cash flow can be divided into the net cash flow and the free cash flow as the future income flow. The net cash flow discounts the cash flow generated by the all business activities in the future. Because business's life could be limitless, the cash flow in the future is viewed as the specific assumption. The net cash flow is also known as ending cash and cash equivalents in the cash flow statements. The free cash flow (FCF) is the remaining cash flow generated after essential capital expenditures with positive net values deducted from the cash flow from operations. The capital expenditures include annual fixed assets expenditures and long-term investment. FCF is usually paid to both the creditors and shareholders.

EVA is not only the indicator of a business creating economic values for shareholders but also the earnings a business creates after the cost of capital deducted from the net operating profit after tax. In other words, when using the fund effectively, a business creates the value higher than this cost of capital. Hence, EVA can be regarded as the values created for shareholders. The cost of capital is the relative cost a business must pay in order to collect the funds that the shareholders and creditors provide. The capital structure of a business is composed of common stock, preferred stock, liabilities and convertible debt, which can be used for calculating the weighted average cost of capital. Generally speaking, obtaining business values is the invested capital added to the anticipated current value of EVA.

3.5.1.2. Market-based approach. The market-based approach is a measure indicator according to the open listed company's price, assets, and earnings information. The value indicator is adjusted according to the appraised company's operating revenue, and then its value is inferred due to the adjusted value indicator. The value indicator is usually the stock price of the listed company divided by certain earnings coefficients, for instance, net sales revenue, net income after tax, pretax net income, net cash flow, pretax cash flow, dividend distribution, and some assets such as the book value of assets, the book value of physical assets, and the book value of the physical assets after adjusting.

3.5.1.3. Asset-based approach. The asset-based approach is adjusted for the balance sheet of a business including tangible, intangible assets, and contingent assets or liabilities to be adjusted according to the fair market price, the replacement cost, or clear accounts. The adjusted assets and liabilities are subtractive and then the adjusted business value is received. The adjusted business value is adjusted again according to the company with control powers and stock rights with its public circulation. If the valuation target is for estimating a few stock right values or an unlisted small business, its business value must be changed downwards in accordance with its control power and market circulation. Finally, its rational business value is received.

The valuation of assets and liabilities are different in accordance with valuation targets. Hence, the fair market price of a going-concern business must be received. If it is difficult to receive fair market prices or its non-existed assets and debt is not in the market, other just valuations or the similar replacement costs can be viewed as the business value. For instance, the real estate depending on the appraiser's valued results or intangible assets in accordance with their replacement costs can be adjusted. If the business is liquidated or will be nonexistent in the future, the proper liquidation method should be found to obtain the liquidation values.

The asset-based approach must be valuated to the assets and liabilities. The accounts include: current assets, physical assets, real estate, intangible assets, short-term liabilities, long-term liabilities, contingent liability, and other special liabilities.

3.5.1.4. Conceptual hierarchy. The design of the valuation model knowledge base can illustrate the concept hierarchy relationship among the financial data, models and computed parameters as illustrated in Fig. 8. The design objective can make users and financial experts understand the hierarchical relationships between Account Names and valuation models from financial statements, notes to financial statements, and financial news. The knowledge base contents include the hierarchical relationships among accounts, valuation models, and computed parameters. The specific concept is not only shown in such a hierarchical relationship, but the unnecessary data can also be filtered out so as to maintain certain data quality. We develop the valuation model knowledge base based on the scope with the valuation and the tree-based concept structure generated. However, all models and parameters in three Business Valuation Approaches are not exactly constructed. In the structure, we view these nodes as essential parameters in valuation models. Its terminal nodes are viewed as Account Names or their synonyms, and the non-terminal nodes are viewed as

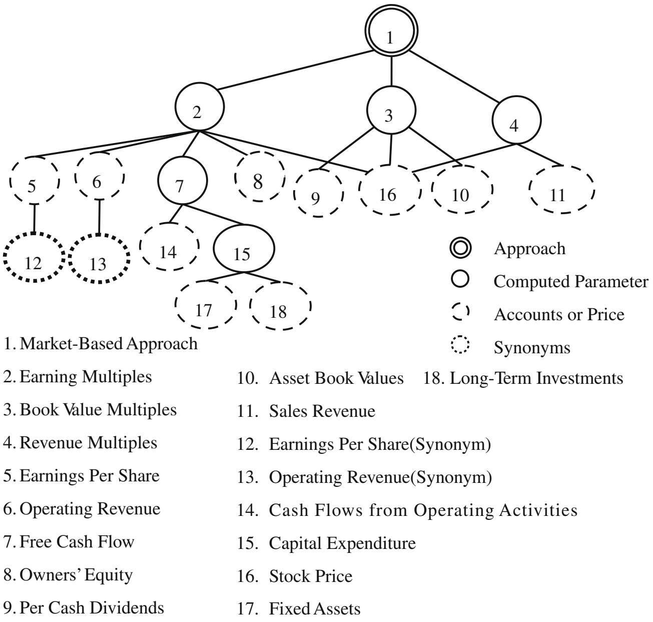


Fig. 8. Concept hierarchy of business valuation.

computed parameters by valuation model. The root is the valuation approach names. In other words, the child nodes are derived from the father node of approach.

In the concept tree, we view the concept hierarchy as a hierarchy tree. A high-level conceptual entity is stored in the place closest to the root (Han, Cai, & Cercone, 1993). On the contrary, the concrete entity is stored in the place closest to the leaf node. The situation is called materialization between two entity types (Goldstein & Storey, 1994). Hence, we can understand that the concept in the asset-based approach is included in the concept of business valuation. Even the two concepts are viewed as an inheritable relationship. For example, the articles in business valuation are not sure to be mentioned about a concept of the assets-based approach but the articles about the assets-based approach must be mentioned. We find out that the materialization includes the existence of inheritable relationships.

The research model shows how to develop information extraction method for the business valuation through Intelligent word segmentation, keyword analysis, extraction techniques, and the knowledge base design with a hierarchy concept showing hierarchical relationships among valuation models from financial statements, notes to financial statements, and financial news. The design objective can precisely extract the necessary financial data for business valuation and the existent relationships between the extracted Account Names and valuation models. Then, we regard the research model as the main research structure and develop it into the physical prototype system to prove the feasibility and validity of the research model.

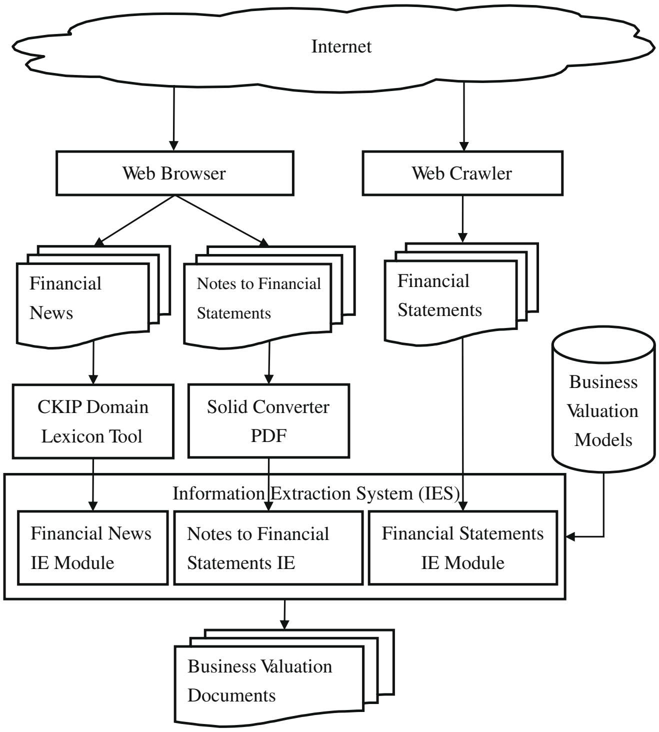

Internet
Web Browser
Web Crawler
Financial News
Notes to Financial Statements
Financial Statements
Business Valuation Models
CKIP Domain Lexicon Tool
Solid Converter PDF
Information Extraction System (IES)
Financial News IE Module
Notes to Financial Statements IE
Financial Statements IE Module
Business Valuation Documents

Fig. 9. Prototype system structure.

4. Prototype system development

In this section, we describe a prototype system according to the research model which is shown in Fig. 9. First, we discuss the system structure and design in the prototype development and illustrate information extraction system functions which meet the design specification.

4.1. System design

4.1.1. Web crawler design

Web crawler is the search technique by which we gather web pages from the Internet for the primary purpose of indexing and supporting a search engine. The objective of crawling is to quickly and efficiently gather as many useful web pages as possible, together with the link structure that interconnects them (Manning, Raghavan, & Schutze, 2007). Hence, in order to save time in data source extraction, we develop a simple web crawler to extract specific financial statement web pages in and store these HTML files on the web server.

4.1.2. Domain Lexicon Tool

The Domain Lexicon Tool developed by CKIP of Academia Sinica is the Chinese word segmentation tool. This tool nearly includes over 100,000 Chinese word lexicons at present. These Chinese lexicons include common lexicons, common named entities, idioms, proper lexicons in a few special fields. Except for the lexicons collected by different fields, the tool offers users a self-constructed dictionary function. We can understand each lexicon meaning in a document. It is to make the prototype system extract and parse correctly each lexicon we need according to its POS tagging. Due to having a self-constructed dictionary function in this tool, we can set up a Chinese Account Name dictionary and a Chinese organization name dictionary, respectively.

4.1.3. PDF conversion tool

Except for data sources such as financial statements and financial news in this research being stored in web pages, the annual reports are stored in PDF files. In order to extract data from PDF files, we use Solid Converter PDF tool to convert PDF files into TXT files. However, during the process of converting, especially when detecting tables, there can be some problems in header position recognition due to table structures in the original documents being special. These errors cause the data records to be missing or incomplete. In other words, when wanting to extract a complete table, programs must first detect the correct header position and then recognize the whole table structure so as to extract the raw data in the table. Because the technique of detecting tables is used in image processing, we cannot make any discussion or propose any solutions. What we is dealing with is how to extract table data generated by the converting tool.

4.1.4. Knowledge base design

We use the enhanced entity relationship (EER) data model to show the knowledge base structure. The EER notation allows us to use the concept of generalization and specialization with attribute inheritance techniques that is applied into business rules. These two techniques assist users to identify top-down and bottom-up relationships. In Fig. 10, entities are shown by rectangles and relationships are shown by diamonds. The knowledge base is composed of four entities. These entities include the Account Name, Business Valuation Model, Business Valuation Approach, and Account Synonym.

Account Names, Business Valuation Approach, and Business Valuation Model are three main strong entities in the knowledge base. We define a supertype called Business Valuation Approach, with subtypes for Business Valuation Model and Account Name. The double-line extended from the Business Valuation Approach entity type to the circle specifies each entity instance of Business Valuation Approach which must be a member of either Business

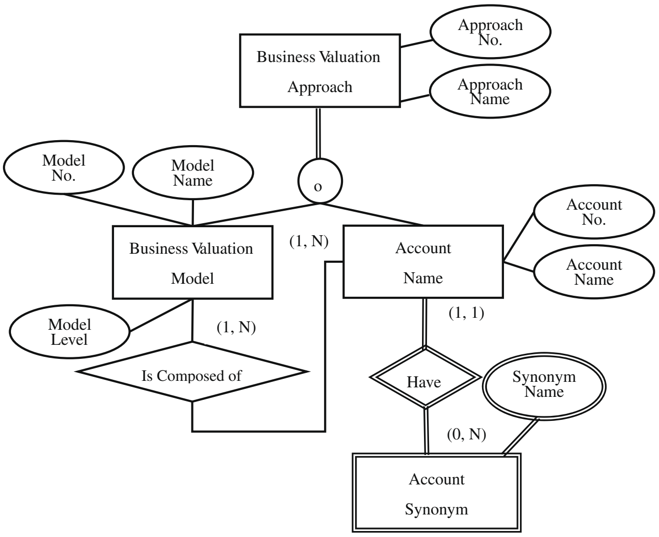


Fig. 10. EER data model for valuation model knowledge base.

Valuation Model or Account Name and must be defined as the total specialization rule. The letter "o" in the circle joining the supertype and subtype specifics an entity instance which is a member of two (or more) subtypes. The situation is called overlap rule.

Account Name is composed of two attributes, Account No. and Account Name. Business Valuation Approach is composed of two attributes, Approach No. and Approach Name. Business Valuation Model is composed of three attributes Model No., Model Name and Model Level. Model Level is represented as the level number of models or parameters in the concept hierarchy. Account Synonym shown by double-line rectangles is a weak entity and is composed of Synonym Name. Such entities are those that cannot exist unless another entity also exists. Hence, the existence of Account Synonym must depend on Account Name. The relationship between a weak entity and it owner is called an identifying relationship which is represented as double-line.

4.1.5. Module design

The information extraction system (IES) is divided into three main functions as shown in Fig. 11. These functions include the user interface, the Analysis Lexicon module, and the information extraction module. We input the segmented financial news, financial statements, and note to financial statements to be the test cases. The proper modules are selected from different data sources and output the experiment results to be stored in a document.

The function of Analysis Lexicon modules searches for continuous segmented words with named entities morphemes to merge a complete lexicon and to change their POS tags by four named entity analysis methods in financial news. These four analysis modules are composed of Organization Name Analysis Module, Account Name Module, Time Module, and Money and Percent Module. These design functions are made according to keyword analysis

The information extraction modules extract specific lexicons to collect business valuation data according to different data sources financial statements, notes to financial statements, and financial news by different extraction algorithms. At this moment, these extracted Account Names are viewed as input parameters put into the specific function to search for relevant valuation model names in the knowledge base to store the business valuation documents.

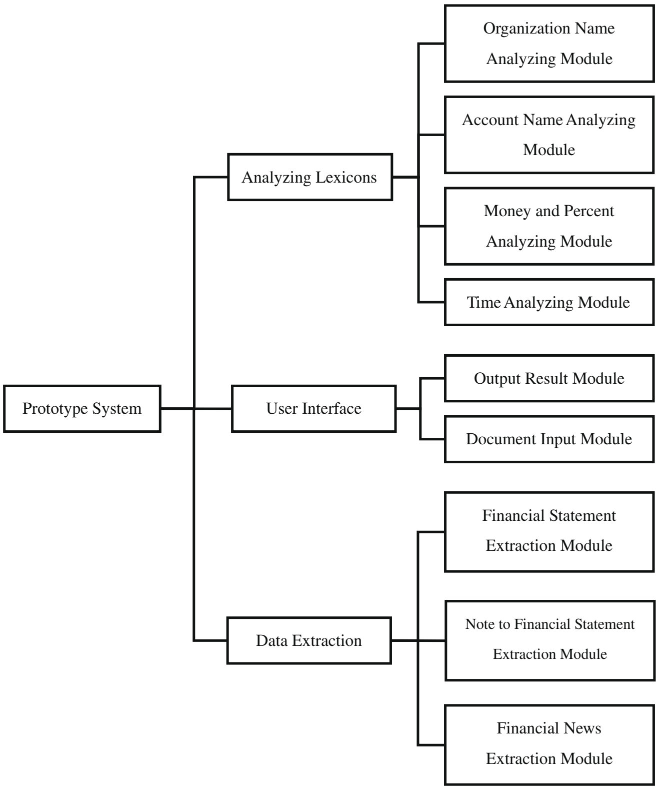

Prototype System
Analyzing Lexicons
Organization Name Analyzing Module
Account Name Analyzing Module
Money and Percent Analyzing Module
Time Analyzing Module
User Interface
Output Result Module
Document Input Module
Data Extraction
Financial Statement Extraction Module
Note to Financial Statement Extraction Module
Financial News Extraction Module

Fig. 11. Prototype system modules.

5. Research experiment

5.1. Experiment design

This research objective is to search for and extract the financial data for business valuation from three data sources, namely financial statements, notes to financial statements, and financial news. The experiment design refers to the research model to be performed. Once we develop the model into the physical prototype system, the system can search for financial data for business valuation in the specific data sources. This allows both experts and normal users to extract correct information from a large amount of data in a short time.

The prototype system extracts financial data for business valuation from the specific data sources. Hence, we must understand the performance of the research model from extraction results. In order to evaluate performance of the model, we adopt benchmark indicators commonly used in information extraction to evaluate the research model, namely, the precision rate and the recall rate.

$$Recall = \frac{\text{the number of extraction words correctly recovered}}{\text{the number of words}}$$

$$Precision = \frac{\text{the number of extraction words correctly recovered}}{\text{the number of extraction words}}$$

In above-mentioned formula, recall rate is the ratio of the number of extraction words correctly recovered to all words of documents. Precision is the ratio of the number of extraction words correctly recovered to the number of extraction words.

We introduce three experiments according to different data sources in the section. The first experiment uses the financial statements in annual reports. The second experiment uses notes to financial statements in annual reports. The third experiment uses the financial e-news from China Times web site. All experimental samples are used to evaluate and perform the experiment design. We illustrate the experimental process and results below.

5.1.1. Financial statements experiment

Experiment I used financial statements as the experimental samples, specifically annual reports issued by listed companies in 2006. We selected a total of 200 financial statements including balance sheets, income statements, cash flow statements, and owner's equity statements from the top 50 listed electronics companies based on owner's equities in 2006. However, we found that balance sheets and income statements of two companies have not been issued. We deducted four from 200 samples so the actual experimental samples total 196.

After loading the sample into the information extraction system as shown in Fig. 12, we chose "balance sheet" as the document type to extract financial data for business valuation to be stored in the output document. We used the experiment results with financial data for business valuation to evaluate the recall rates and the precision rates.

In experiment I, we have four experimental results divided according to statement sources. In the experimental results, because every sample layout is almost the same on web pages, the specific time and organization names can be easily extracted in four statements. But, we are concerned about whether the Account Names and money can be extracted in statements. In the experiment results, we find that both the recall rate and the precision rate get up to 100% for the income statements because the Account Names used on these statements are almost all correctly identified. We also find that synonyms do not appear on the income statements. Although the other three statements have good recall rates and precision rates, the system is sometimes unable to recognize all so as to be ignored in the appearance of the synonym word. For recall rates, synonymous words in the balance sheets are more conspicuous than in the owner's equity statements and the cash flow statements. For precision rates, keywords extracted in the cash flow statements are usually incorrect, which in turn generates lower precision rates. As for computer run time, the average run time of every test sample was 0.6 s. We know that the retrieval speed for computers is faster than when done by hand (see Fig. 13).

5.1.2. Notes to financial statements experiments

Experiment II used annual reports as experimental samples. We selected 137 table samples of notes to financial statements in annual reports from the top 50 listed electronics companies according to company capital in 2006.

When converting the table of notes to financial statements in the PDF file into the EXCEL file, we used the PDF converting tool first and chose the file we wanted to convert. We used experiment results with financial data for business valuation to evaluate the recall rates and the precision rates.

In experiment II, we used the tables in notes to financial statements as the samples because there are contents of significant accounts which are regarded as essential information for business valuation in these tables. In the experimental results, there are 123 samples (tabular data) associated with business valuation as shown in Fig. 14. Money and percent scored nearly 100% on the recall rate and precision rate because they are different from the general Chinese words and the system is able to extract them easily. Time is less apparent than the other two named entities in all experimental samples. Though the recall rate of time is more than 90%, the precision rate is very low. Because we used the time morpheme method to search for specific time in the test samples, the specific time we searched for was apt to be confused.

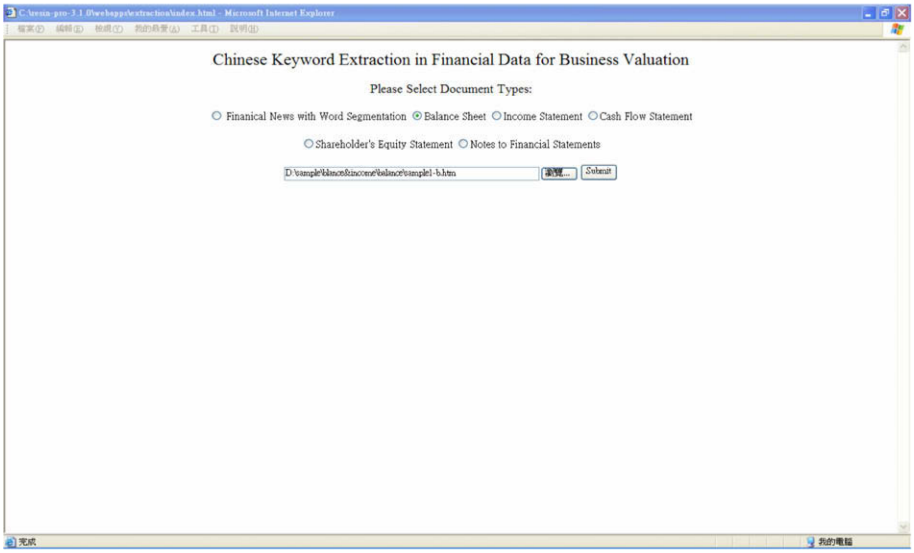
5.1.3. Financial news experiments

Experiment III used China Times financial e-news. We selected 200 pieces e-news of the top 50 companies tested above in one month as shown in Fig. 15. Because e-news web pages are stored in HTML and XHTML documents, we must get rid of web page tags so as to convert to a XML file compatible with the segmentation tool and then load the XML file into the Domain Lexicon Tool to resolve lexicons and POS tags.

After the word segmentation process, we took the word segmentation contents in the samples and then loaded the contents into the keyword analysis and information extraction system. We chose "Financial News with Word Segmentation" as the document type to extract financial data for business valuation to be stored in the output document.

From experiment III results, we can understand that the precision rate (100.00%) and the recall rate (99.38%) of Account Names have the best results in four keyword methods designed because we used the self-constructed dictionary function to make the word segmentation process on the Account Names.

Regarding extraction of business valuation data, we extracted financial data according to the information extraction method. The system extracted 541 sets of financial data in all experimental


C:\resin-pro-3.1.0\webapps\extraction\index.html - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
瀏覽...
Submit
完成
我的電腦

Chinese Keyword Extraction in Financial Data for Business Valuation

Please Select Document Types:

○ Finanical News with Word Segmentation ● Balance Sheet ○ Income Statement ○ Cash Flow Statement

○ Shareholder's Equity Statement ○ Notes to Financial Statements

D:\sample\blance&income\balance\sample1-b.htm

Fig. 12. Information extraction on financial statements.

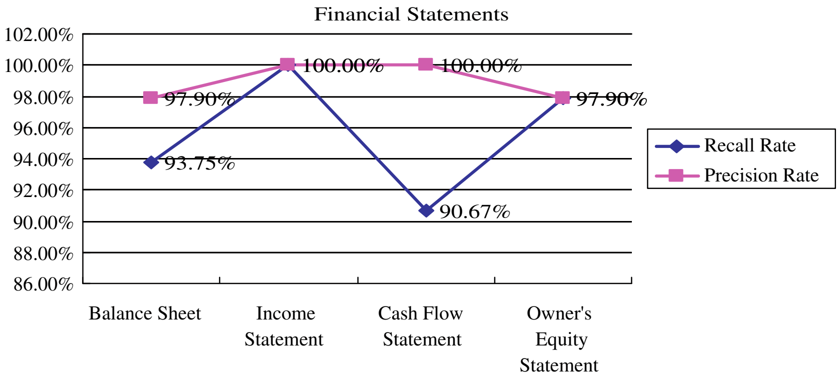


Fig. 13. Experiment results on financial statements.

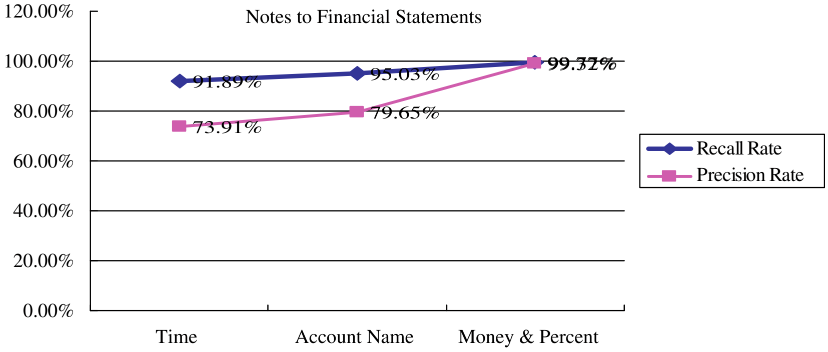
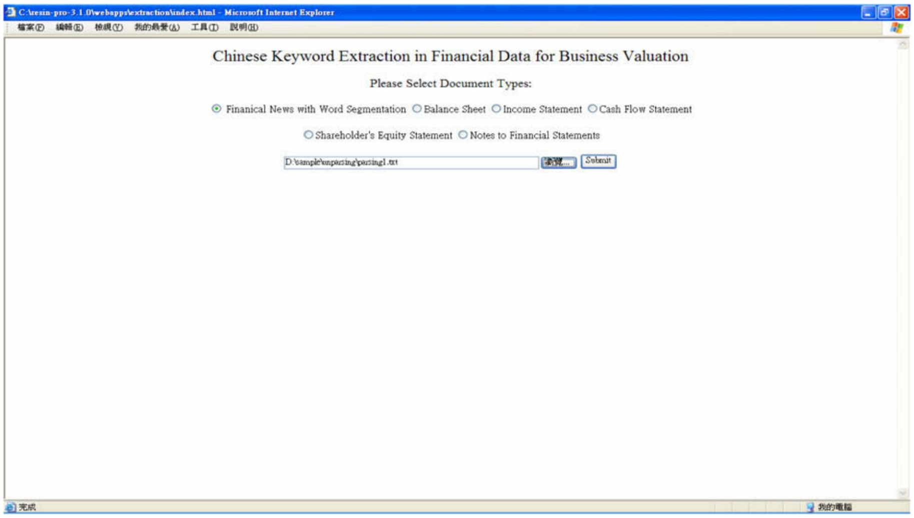


Fig. 14. Experiment results on notes to financial statements.


C:\resin-pro-3.1.0\webapps\extraction\index.html - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
瀏覽...
Submit
完成
我的電腦

Chinese Keyword Extraction in Financial Data for Business Valuation

Please Select Document Types:

● Finanical News with Word Segmentation ○ Balance Sheet ○ Income Statement ○ Cash Flow Statement

○ Shareholder's Equity Statement ○ Notes to Financial Statements

D:\sample\unparsing\parsing1.txt

Fig. 15. Keyword analysis and information extraction system on financial news.

samples. There were 485 correct and complete business valuation data in the extracted data. There were 535 real data in all samples as shown in Fig. 16. The evaluation method is to determine if each piece of financial data about business valuation is completely and correctly extracted. In the e-news, the system sometimes extracted redundant words which affected experiment results due to complicated Chinese grammar structures. But, if extracted data still had complete meanings as the final evaluation objective, it was extracted. The recall rate and the precision rate are 88.67% and 87.84%, respectively, in the experimental results. In computer run time, the average run time of every test sample is 1.354556 s.

6. Research implication and conclusion

6.1. Research implication

In this section, we discuss the research findings and implications according to the research model and experiment results. We illustrate managerial and technical findings and implications.

Information extraction is a method to search for important specific information or to extract specific relationships among information from documents. We use the specific extraction technology to search for financial data for business valuation from

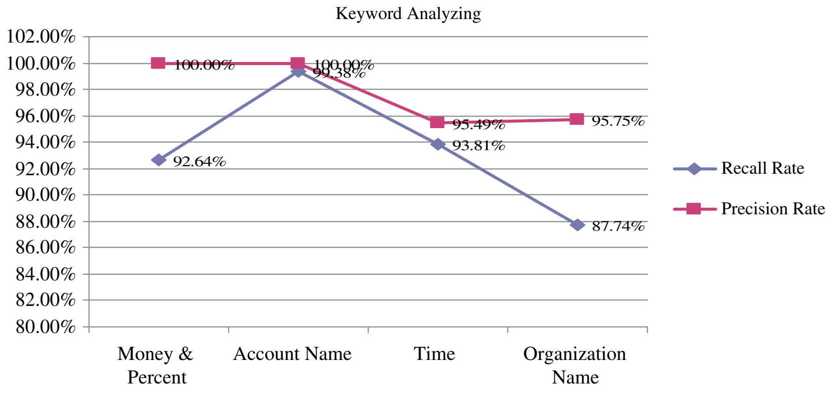


Fig. 16. Experiment results of keyword analysis on financial e-news.

financial statements, notes to financial statements, and financial news. In the past, valuation experts had to collect the raw data on-line by themselves and read and filtered the necessary information in detail by manual methods. Therefore, they often had to spend a large amount of time searching for information and there were no relevant specific web sites or databases to make inquiries and analyze data about business valuation. The research results show that we can extract nearly correct valuation data from a large number of financial statements, notes to financial statements, and financial news to improve the exactness and efficiency of text processing.

In the literature, authors were in favor of describing and introducing specific Business Valuation Models instead of discussing relationships among models. We attempt to design and complete the conceptual hierarchy of business valuation in accordance with financial parameters. When we can extract financial data from these data sources, the relationship between the Account Name and the Business Valuation Models is revealed. This allows users, valuation industries and financial experts to prove and quickly extract the existing relationships from the specific data sources. The existent relationships makes users analyze, verify, and calculate correctly these data. We also find out that the conceptual hierarchy in business valuation is able to show specific hierarchical relationships. The hierarchical relationships cannot only express the specific concepts among valuation models, but also filter out useless data not relevant business valuation so as to maintain and verify the data quality.

Information extraction must be achieved by three methods. They are the word segmentation, syntax analysis, and information extraction in pure-text documents with unstructured data. However, Intelligent word segmentation still has incorrect segmentation results which create the possibility for incomplete or lost data. In order to solve these problems, we first define POS tags of keywords. We extract for business valuation data and analyze the keywords. In keyword extraction, we extract pre-defined data according to new POS tags and extraction rules.

In the keyword analysis, we find out few studies carried out the Chinese abbreviations of organization names in word segmentation field. In this research, we extract the longest common subsequence between the named entity dictionary and word segmentation and show the recall rate and precision rate achieving a high-level performance via the experimental results. Account Names using the dictionary method achieves 100% precision rate. Time, money, and percent using POS tagging and morpheme analysis determine the continuous segmented word sequences and then form one complete noun. This method achieves more than the precision rate and the recall rate of 90%. Hence, this research not only improves performance in named entities recognition but is also widely applicable to the extraction of named entities.

In the financial statement extraction processing, we find out that regular expressions can extract raw data in specific field almost exactly on web pages. In notes to financial statements processing, due to converting tools not achieving complete output data, the evaluation results are still affected in named entities recognition. But the findings indicate that money and percent still have good performance. In financial news processing, we use defined POS tags and extraction rules to extract and to collect the valuation data to be stored in a document. The experiment results show the extraction method is reliable. Overall, the findings suggest that the extension of Intelligent information extraction in business valuation achieves conspicuous performance. In other words, this research has taken a step in the application of information extraction.

6.2. Conclusion and future research work

We extend information extraction techniques in business valuation. We develop and design extraction methods in financial data for business valuation from different data sources. We can see that the research model achieves conspicuous performance in extraction methods from different data sources. If business valuation databases are developed in the future as in foreign countries, we will collect correct and complete raw data from the heterogeneous data sources. We not only save a large amount of time on the process, but also can take correct business valuation data quickly from these data sources.

Although this research can automatically extract business valuation data from different heterogeneous data sources, we still propose some problems to improve the research model. We offer suggestions for the future research.

- *Extraction design problems:* In the process of extracting financial data, Chinese article structure, syntax, and meaning lacking fullness and exactness, sometimes cause the system to be apt to extract incomplete or incorrect data. These errors are unable to express complete meanings of documents and cause misunderstanding.
- *Extraction methods for different data sources:* In addition to the data sources we use, there are different data sources, for example, annual reports, prospectus, industry reports, and trade union data, which also include the important information about business valuation. We need to extract the extraction technique in the future to apply to these heterogeneous data sources.
- *Integral business valuation database:* In order to fulfill the growing and changing business valuation data needs, if integrating and developing with the existing heterogeneous data sources, we cannot only offer effective references for corporate governance supervision, information disclosure, and academic

research, but also offer these as important indicators for evaluating business activities when in public listing and in merge and acquisition.

Acknowledgment

This research is sponsored by National Science Council Research Grant No. NSC95-2416-H-004-006.

References

Atlam, El-S., Fuketa, M., Kashiji, S., Nakata, H., & Aoe, J. (2002). A new method for construction filed association terms using co-occurrence words and declinable words information. *IEEE International Conference on Systems, Man and Cybernetics, 4*, 1217–1224.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Addison Wesley Longman Publishing Co..

Cercone, N., Huang, X., Peng, F., & Schurmans, D. (2003). Applying machine learning to text segmentation for information retrieval. *Information Retrieval, 6*(3), 333–362.

Chen, K. J., & Ma, W. Y. (2001). Construction and management for Chinese corpus. In *Proceedings of the research on computational linguistics conference* (pp. 175–191).

Chen, A., Gey, F. C., He, J., Meggs, J., & Xu, L. (1997). Chinese text retrieval without using a dictionary. In *Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 42–49).

Chen, K. J., & Ma, W. Y. (2005). Design of CKIP Chinese word segmentation system. *Chinese and Oriental Languages Information Processing Society, 14*(3), 235–249.

Chien, L. F., & Pu, H. T. (1996). Important issues on Chinese retrieval. *Computational Linguistics and Chinese Language Processing, 1*(1), 205–221.

Goldstein, R. C., & Storey, V. C. (1994). Materialization. *IEEE Transactions on Knowledge and Data Engineering, 6*(5), 835–842.

Han, J., Cai, Y., & Cercone, N. (1993). Data-driven discovery of quantitative rules in relation databases. *IEEE Transactions on Knowledge and Data Engineering, 5*(1), 29–40.

Krupl, B., Herzog, M., & Gatterbauer, W. (2005). Using visual cues for extraction of tabular data from arbitrary HTML documents. In *Special interest tracks and posters of the 14th international conference on World Wide Web* (pp. 1000–1001).

Li, W., Wong, K. F., & Yuan, C. (2003). A design of temporal event extraction from Chinese financial news. *International Journal of Computer Processing of Oriental Languages, 16*(1), 21–39.

Liu, Y., Mitra, P., Giles, C. L., & Bai, K. (2006). Automatic extraction of table metadata from digital documents. In *Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries, JCDL'06*.

Lochovsky, F. H., & Wang, J. (2003). Data extraction and label assignment for Web database. In *Proceedings of the 12th international conference on World Wide Web* (pp. 187–196).

Manning, C. D., Raghavan, P., & Schutze, H. (2007). *An introduction to information retrieval*. Cambridge, England: Cambridge University Press.

Rosenfeld, B., Feldman, R., & Aumann, Y. (2002). Structural extraction from visual layout of documents. In *Proceedings of the 11th international conference on information and knowledge management* (pp. 203–210).

Zhai, Y., & Liu, B. (2005). Web data extraction based on partial tree alignment. In *Proceedings of the 14th international conference on World Wide Web* (pp. 76–85).