An Automatic Indexing and Neural Network Approach to Concept Retrieval and Classification of Multilingual (Chinese-English) Documents

Chung-hsin Lin and Hsinchun Chen

Abstract—An automatic indexing and concept classification approach to a multilingual (Chinese and English) bibliographic database is presented. We introduced a multi-linear term-phrasing technique to extract concept descriptors (terms or keywords) from a Chinese-English bibliographic database. A concept space of related descriptors was then generated using a co-occurrence analysis technique. Like a man-made thesaurus, the system-generated concept space can be used to generate additional semantically-relevant terms for search. For concept classification and clustering, a variant of a Hopfield neural network was developed to cluster similar concept descriptors and to generate a small number of concept groups to represent (summarize) the subject matter of the database. The concept space approach to information classification and retrieval has been adopted by the authors in other scientific databases and business applications, but multilingual information retrieval presents a unique challenge. This research reports our experiment on multilingual databases.

Our system was initially developed in the MS-DOS environment, running ETEN Chinese operating system. For performance reasons, it was then tested on a UNIX-based system. Due to the unique ideographic nature of the Chinese language, a Chinese term-phrase indexing paradigm considering the ideographic characteristics of Chinese was developed as a multilingual information classification model. By applying the neural network based concept classification technique, the model presents a novel way of organizing unstructured multilingual information.

I. INTRODUCTION

The overwhelming volume of online information generated and disseminated across computer networks has created a significant burden for researchers and practitioners. For structured numeric data, database management systems (DBMS) have typically been used. However, for unstructured textual data, information management, processing and retrieval remain very complex and problematic.

Information Retrieval (IR) is a research area that has been studied extensively in the western world [8], [43]. For example, retrieving information from large, unstructured databases of English documents has been an area of inquiry for several decades. Nevertheless, due to the cognitive process involved in organizing and retrieving information [8], there still exists significant uncertainty and difficulty in online information management and information retrieval.

The prevailing keyword searching and user browsing techniques for IR suffer from an inability to support concept or content-based search [2]. Recently, researchers and practitioners have turned their attention to "concept-based" IR, e.g., the (automatic) concept space approach proposed by Chen and his co-workers [10], [11] and the (manual) thesaurus search and content-based IR supported by commercial packages such as Verity's Topic and Oracle's ConText. Such a concept-based retrieval capability has been considered by many researchers and practitioners to be an effective complement to the prevailing keyword search or user browsing options available in most information retrieval systems.

For oriental languages, the situation is even more complex than for western languages [C10]. Due to the unique linguistics and grammar structures of oriental languages, IR techniques applicable to western languages may not be appropriate for oriental language information organization and retrieval. Although several techniques have been proposed, an automatic and systematic approach to multilingual information management and retrieval is still lacking. In this research, we examined a multilingual bibliographic database which contained technical documents mainly written in Chinese, with occasional English terms. This application allowed us to study some crucial multilingual information retrieval issues and helped us develop an integrated, concept-based Chinese-English information classification and retrieval model.

The structure of this article is as follows. Section II discusses information retrieval problems in the context of multilingual databases and presents our proposed concept space approach. Extensive literature reviews are also provided. Section III presents a classical English classification model and our proposed multilingual model. Section IV describes in detail the automatic multilingual indexing techniques adopted in this research, in particular for Chinese information. Section V presents a neural network based concept classification technique. Based on cluster analysis and Hopfield network algorithms, an unstructured database of multilingual documents can be organized in a semantically structured form. Section VI summarizes a system implementation for a sample...
bibliographic database of 1052 documents. Conclusions and
directions for additional research appear in Section VII.

II. PROBLEM DESCRIPTION AND RESEARCH DESIGN

Different human languages exhibit significantly different
linguistic and grammatical characteristics which strongly affect
how information is structured and represented in modern
databases. This is particularly true of the contrasts between
western languages (e.g., English, French, German, etc.) and
oriental languages (e.g., Chinese, Japanese, Korean, etc.).
Despite these differences, there are common problems associ-
ated with online information management and retrieval across
databases created in different languages. In this research we
attempted to address the information management and retrieval
issues related to a multilingual database containing mainly
Chinese and English texts.

In this section, we will first review some common IR prob-
lems and the prevailing techniques to address these problems.
Discussion of specific research objectives and the techniques
adopted in this research will follow.

A. Information Retrieval Problems and
Techniques: An Overview

In the past few decades, the availability of cheap and effec-
tive storage devices and information systems has prompted
the rapid growth and proliferation of relational, graphical,
and textual databases. Information collection and storage have
become easier, but effort required to retrieve relevant informa-
tion has become significantly greater, especially in large-scale
databases. This situation is particularly evident for textual
databases, which are widely used in traditional library sci-
ence environments, in business applications (e.g., manuals,
newsletters, and electronic data interchanges), and in scientific
applications (e.g., electronic community systems and scientific
databases). Information stored in these databases often has
become voluminous, fragmented, and unstructured after years
of intensive use. Only users with extensive subject area knowl-
edge, system knowledge, and classification scheme knowledge
are able to maneuver and explore in these textual databases [7].

In conventional information retrieval environments, key-
words are manually or automatically assigned and queries
are formulated by using terms interconnected by Boolean
operators. Although widely used, the Boolean query languages
have some drawbacks: Users find it difficult to formulate their
queries using the Boolean syntax, the retrieved documents are
not ranked in any particular order; and most importantly, the
retrieval results are often inadequate [43], [44]. The vocab-
ulary problem in human-computer interactions further
confound the keyword-based Boolean retrieval mechanism [5].
In [23], Furnas et al. found that in spontaneous word choice
for objects in five domains, two people favored the same term
with less than 20% probability. This fundamental property of
language limits the success of various design methodologies
for keyword-driven interaction.

The vector space model, proposed by Salton [43], [44],
provides an alternative approach for handling information re-
trieval applications. In the vector space model, both the stored
documents and the user queries are represented by sets of terms
and weights pairs, without Boolean operators. Documents
and queries are then compared based on similarity functions
(e.g., cosine function, Jaccard’s function, etc.) between the
matching terms [16], [43]. The similarity scores of the relevant
documents indicate a match between a user’s query and the
related documents and can be ranked and displayed to the
users more meaningfully.

Most recent internet resource discovery services support
basic Boolean or vector space querying capabilities. For ex-
ample, Gopher permits keyword-based search across different
information sources and WAIS allows search using vector
space matching. In addition to these keyword-based searching
functionalities, internet resource discovery software often also
provides user browsing capabilities. For example, Gopher
permits its user to traverse in a system hierarchy (a road
map for the Gopher space) and Mosaic supports hypertext
browsing across information sources. However, several design
problems frequently arise from user browsing, especially in a
large information space. The browsing system can potentially
confuse and disorient its user through what is known as the
embedded digression problem and it can cause the user to spend
a great deal of time while learning nothing specific, the
art museum phenomenon [20], [3]. Hypertext systems provide
bookmarks, context maps, backtracking or history lists, and
guided tours to help the user overcome such problems [38].
However, browsing in a large and unfamiliar information space
is still time-consuming, cognitively demanding, and at times
non-productive.

Based on our experiences in dealing with several textual
applications, including intelligence analysis [10], [11], meet-
ing support systems [9], and scientific (molecular biology)
databases [12], we have developed an algorithmic, concept-
based approach to information classification and retrieval. In
our design, we generate a concept space by first extracting
concepts (terms) automatically from the texts in databases.
Similar concepts are then linked through the co-occurrence
analysis of concepts in texts. The concept space created
represents the vocabularies used in documents and the similarity
probabilities between these vocabularies. A graph traversal
and/or clustering algorithm can then be used to help searchers
identify similar concepts in the concept space and “dock” on
to the relevant part of the information space automatically. We
believe the difficulties and problems associated with keyword-
based searching and user browsing can be partially alleviated
by adopting the proposed concept-based IR approach. We
present a blueprint of our approach and a review of relevant
literature below. Details about the specific algorithms will be
discussed in the next section.

• Concept Identification: Despite the increasing availability
of other presentation media such as images, voices, ani-
mations, and videos, the most natural and popular means
of communication is still natural language. In system-
supported applications, large-scale online textual output
could reveal the concepts of a specific domain.

The first task for concept space creation is to identify
the vocabularies used in the textual output. AI-based
natural language processing (NLP) techniques such as
the Augmented Transition Network (ATN) parsing, case grammar, and semantic grammar have been used for creating unambiguous internal representation of English statements. However, because such techniques are either too computationally intensive or are domain-dependent, they are inappropriate for identifying content descriptors (terms, vocabularies) from texts. An alternative method for content identification that is simple and domain-independent is the automatic indexing technique, often used in information science for indexing literature. In [43], Salton presents a blueprint for automatic indexing, which typically includes dictionary look-up, stop-sentence, word stemming, and term-phrase formation. The algorithm first identifies individual words. A stop word list is then used to remove non-semantic bearing words such as the, a, on, in, etc. After removing the stop words, a stemming algorithm is used to identify the word stem for the remaining words. Finally, term-phrase formation that formulates phrases by combining only adjacent words is performed.

- **Linking Similar Concepts**: While automatic indexing identifies vocabularies used by different group members (from the texts), the relative importance of each term for representing the group members’ concepts may vary. That is, some of the vocabularies used may be more important than others in conveying meanings. Salton’s vector space model [43] associates with each term a weight to represent its descriptive power (a measure of importance). Among the many probabilistic techniques that have been developed by various information science researchers, techniques which typically incorporate term frequency and inverse document frequency have been found to be useful [43]. Term frequency refers to the number of times a term appears in a given document. Document frequency refers to the number of documents in the entire database that contain a given term.) The basic rationales underlining these two measures are that: Terms which appear more times in a specific text should be assigned higher weights (term frequency) and terms which appear in fewer texts (the more specific or unique terms) also should have higher weights (inverse document frequency).

Based on cluster analysis, the vector space model could be extended for concept space generation. The first stage in cluster analysis is to convert the raw data (e.g., terms and weights) into a matrix of similarity measures between any pair of terms. The similarity measure computation is mainly based on the probabilities of terms co-occurring in the documents of a database. The probabilistic weights between terms indicate their strength of relevance or association.

- **Concept Space Traversal and Clustering**: When a searcher encounters a retrieval problem (with keyword searching or browsing), it is conceivable that he/she could consult (browse) the concept space and identify other relevant vocabularies for use. This is in fact what professional librarians do when assisting patrons in finding relevant terms using a thesaurus. (Notice that cluster analysis creates similar links, not synonymous links, although in practice, many synonymous terms will have a high similarity probability between them.) An alternative method for traversing the concept space could be based on system-aided, multiple-link searching algorithms [10]. For example, a Hopfield network search could traverse the concept space in a parallel mode and combine evidence from multiple links until the search algorithm converges and finds terms most relevant to a searcher’s initial query [11].

In addition to using a graph traversal algorithm to navigate a semantics-rich concept space (and eventually locate relevant documents in the information space), we can also use selected clustering algorithms to partition the graph and extract a small number of “concept groups” (areas of the graph which contain very strongly related concepts). These concept groups can be used to represent the subject matters of an information source. Representing information sources succinctly and accurately is one of the most difficult challenges in the increasingly popular internet resource discovery services, where tens of thousands of information sources need to be accessed transparently [39]. Conventional approaches for linking similar concepts in IR include: single-link clustering and complete-link clustering [49], [42], [41]. Conceptual clustering and neural networks clustering have also been adopted to reveal a small number of overlapping and strongly-associated concepts [51], [37], [47], [48], [9].

By providing a meaningful and semantics-rich concept space and clusters for the often large and unstructured information space using the co-occurrence analysis and clustering techniques, the proposed approach presents a concept-based retrieval option to users, complementary to the prevailing keyword-based searching and user browsing. However, the feasibility of such an approach in the complex multilingual IR environment remains to be examined.

**B. Multilingual Information Retrieval: Chinese-English Bibliographic Databases**

English is a phonographic language in which almost every word has one or more independent meanings [C3]. However, the structure of the Chinese language is based on pictographs and each Chinese word (character) has a unique meaning. In ancient Chinese literature, writing tended to be more concise, with a single word conveying several meanings. In modern Chinese writing, especially for technical literature, term-phrases which contain more than one word have often been used to express precise meaning. A technical term is seldom represented by a single word. Nevertheless, most existing Chinese information retrieval systems are still based on word indexing [C7], [C9], [C13], [C13], [C14], [C15].

A problem associated with Chinese technical literature indexing is referred to as the word-division problem. Automatically extracting meaningful phrases instead of incidental, meaningless words from Chinese literature remains a challenge for researchers [C4], [C5], [C6], [C8], [C12]. Two approaches to the Chinese word-division problem have been proposed. The first of these is grammar-based [18], [54], [35], [53],
Fig. 1. Examples of Porter’s stemming rules.

Incorporating a syntactical and semantic knowledge base into a computer program will enable it to perform semantic parsing of documents and texts [46], [24]. However, the effectiveness of this approach for large applications has not been fully tested. Developing an effective and robust natural language parser for Chinese documents is still in its infancy.

The other approach is non-grammar-based [C1], [50], [17] and typically uses a large built-in dictionary to help extract phrases automatically from unstructured Chinese information. In [C2], the researcher coupled a phonetic (pinyin) symbol with the original ideographic Chinese entry to help extract phrases from Chinese sentences. This approach represents Chinese information in an English-like structure. A space character is used for separating each pinyin symbol in the phonetic symbol file and an algorithm is used to automatically divide and extract phrases appearing in the pinyin symbol file. This approach is a variant of the term-marking operation described in [C11]. Due to the difficulties in creating large-scale, detailed syntactical and semantic knowledge bases (the grammar-based approach), our research adopted the non-grammar-based approach.

Fig. 2. Sample Chinese-English bibliography records.

The phonographic nature of the English language has created the well-documented suffix-stripping problem during information retrieval [40], [34], [26]. It is not clear whether the Chinese language is immune from this problem or not. For example, should “書本” (book) be suffix-stripped to be “書” since these two terms have the same meaning? This research allowed us to examine a few well-known Chinese-English document retrieval problems such as the English suffixing problem. The specific English suffix-stripping algorithm adopted in our research was initially developed by Porter [40]. Some sample rules are shown in Fig. 1.

After addressing the Chinese word-division and the English suffix-stripping problems, we then were able to adopt automatic indexing techniques, the vector space model, and concept space generation techniques [43], [9] to analyze and classify Chinese-English documents. The specific goals of this research included:

- Developing an approach to automatically dividing and extracting term-phrases from Chinese bibliographic documents.
- Developing an integrated, automatic Chinese-English indexing and classification model.
- Generating a concept space and concept clusters automatically to assist in concept-based information retrieval and classification for multilingual databases.

We aimed to improve Chinese information retrieval from the character-word level to the term-phrase level, characteristic of English information retrieval, but still preserve the ideographic features of the Chinese language and then to facilitate automatic extraction and classification of concepts hidden in unstructured multilingual databases.

The test bed for our research was a Chinese-English bibliographic database, initially stored in the MS-DOS environment. Due to the requirement of processing Chinese information, the ENET Chinese operating system (one of the most popular Chinese operation systems on PCs) was also used. Our prototype system was developed in ANSI-C. The system was then ported to a UNIX SPARC 390 workstation for systematic analysis and testing.

The sample bibliographic database contained a collection of articles published in ENET magazine over the past four years. Fig. 2 presents sample entries. Each entry consists of structured data (i.e., the author, the publication year, and the volume number) and unstructured data (i.e., the title). Many titles contain both Chinese and English terms, a characteristic of technical Chinese databases. In the multilingual system model proposed in this paper, we were concerned with only
III. A MULTILINGUAL INFORMATION CLASSIFICATION MODEL

A. An English Classification Model

The stages involved in indexing and classifying unstructured English documents according to the concept space approach proposed by Chen and his co-workers can be summarized in the English classification model shown in Fig. 3. In this model, four steps are necessary to automatically extract and represent concept descriptors (terms or keywords).

- The first step, source entry reformation, identifies information units in documents. A sentence, a paragraph, a chapter, or the whole document could be defined as a basic information unit. A spelling correction component [15] can be used to correct misspelled words. Each information unit can then be examined to remove stopwords (words without specific meaning). After stopwording, each remaining word then needs to be "stemmed" (i.e., the root form of a word must be identified.) For example, "compute" is the stem for "computing," "computers," "computation," etc.
- Term-phrase formation forms phrases from adjacent words of each information unit. Due to the phonographic features of the English language, adjacent words extracted from a phrase segment often have their own precise meaning. For example, "Information Retrieval Systems" can form term-phrases like "Information Retrieval" or "Retrieval Systems." Each of these terms preserves a unique semantic. By employing this term-phrase formation step, English documents can be indexed at the term-phrase level instead of the character-word level employed for present Chinese textual databases [52].
- Term-phrase formation generates a collection of concept descriptors to represent the source document. Each descriptor can then be weighted according to term frequency and document frequency [43] and the relationship between any two descriptors can be computed based on term co-occurrence in documents [16], [9]. This step is referred to as relevance weighting in this research.
- The final step, classification, clusters highly relevant concept descriptors into different concept groups [9]. This splits off the complete concept space of descriptors into multiple concept partitions. Thus the original unstructured information can be reconstructed in a conceptually structured and meaningful way. Statistical analysis and neural network techniques have been used for this purpose [43], [16], [9].

B. A Multilingual Classification Model

As shown in Fig. 4, our multilingual classification model consists of six phases, the first four for automatic multilingual indexing and the last two for automatic concept classification. Due to the difficulty in translating foreign technical and scientific terminologies into Chinese, it is a common practice to use both Chinese and foreign terminologies in Chinese technical literature. This phenomenon is evident in our sample bibliographic database of computing related documents. In order to support concept-based management and retrieval of Chinese-English documents, we developed an integrated multilingual information classification model. An overview of the model is presented below. Details about specific techniques are presented in Sections IV and V.

- Source entry reformation re-structured the source entry into a collection of text segments. Stopword checking and stemming were performed for English entries. For Chinese information, a space was added to separate the Chinese words.
- A multi-linear term-phrasing technique was then used to extract term-phrases from the text segments derived
above. In our system, 1-word, 2-word, 3-word, and 4-word phrases were extracted to represent the source entry.

- **Legality checking** was then performed to screen out illegal Chinese phrases extracted in the second phase. This phase is designed to eliminate the *shadow effect* (discussed in Section IV-C) created by the multi-linear term-phrasing phase. Since English terms had been checked already, they were not examined in this phase.

- The fourth phase, **filtering**, performed the equivalent of English stopwording for Chinese term-phrases. A term with no specified meaning should not be used as a concept descriptor. In our system, a domain-specific "stopterm" filter, unlike the meanin- us filter knowledge base in [1], was used to perform the stopterm checking function. Both filters essentially accomplished the same function by filtering out non-semantic-bearing terms and retaining meaningful concept descriptors. Fig. 5 shows a portion of the domain-specific Chinese stopterm list used in our system.

- The fifth phase, **relevance weighting**, applied techniques based on the vector space model and co-occurrence analysis, similar to those used in the English classification model, to calculate the relevance weight between any pairs of concept descriptors. In our system, the algorithm adopted in [9] for electronic meeting comments classification was modified and integrated into the multilingual information classification model.

- In the **parallel classification** phase, we developed a variant of the Hopfield neural network to partition the concept space derived from these five phases. Classification was performed in a parallel, feed-forward manner. Relevant concept descriptors were identified and grouped together.

### IV. Automatic Multilingual Indexing

This section will describe in detail the first four phases of the multilingual classification model. They all contribute to our efforts to resolve the automatic Chinese-English indexing problem. Examples and analyses are presented.

#### A. Source Entry Reformulation

Source entry reformulation reformulates each source entry into a collection of text segments without stopwords. Since there are two languages involved, Chinese and English, both need to be processed.

Each Chinese word was separated from the others by the space character. For technical Chinese literature, a Chinese word often plays the role of a character in English. Therefore, the basic semantic unit in modern Chinese is a term-phrase, not a single word. Words need to be combined in order to obtain specific meanings.

English words in the source entry were checked against a stopword list of 150 entries. An identified stopword was replaced by a `'-' symbol. Remaining words were then processed by the Porter's stemming algorithm [40]. After the stopwording and stemming process, English entries were ready for automatic indexing. However, Chinese entries still needed to be further analyzed.

#### B. Multi-Linear Term-Phrasing

Two approaches were considered for forming phrases (see Fig. 6). One method is called single-linear term-phrasing. No two terms will use the same word from a specified text context. For example, terms that can be extracted from the text segment "Information Retrieval System" include ("Information", "Retrieval", and "Systems"). ("Information" and "Retrieval Systems"), or ("Information Retrieval" and "Systems"). "Information Retrieval" and "Retrieval Systems" will not be generated as a set of candidate terms using single-linear term-phrasing.

We believe a multi-linear term-phrasing approach is more appropriate for extracting complete, meaningful Chinese phrases. This approach can also be adopted for English documents. Extracted phrases could have the same word constituent. For example, for "Information Retrieval Systems", "Information Retrieval" and "Retrieval Systems" can be formed. We found the multi-linear method to produce more phrases than the single-linear method. Both Chinese and English sentences were transformed to a collection of multiple-word terms in the same way.

Fig. 6 shows the different possible permutations of adjacent words which can be generated for the same text segment using the two approaches. Table I presents a summary of the results of multi-linear term-phrasing for the bibliographic database;

---

**TABLE I**

<table>
<thead>
<tr>
<th>Terms</th>
<th>1-word</th>
<th>2-word</th>
<th>3-word</th>
<th>4-word</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terms</td>
<td>15442</td>
<td>12580</td>
<td>9947</td>
<td>7481</td>
<td>45450</td>
</tr>
</tbody>
</table>

---
(1) Single-Linear Term-Phrasing

<table>
<thead>
<tr>
<th>木</th>
<th>上</th>
<th>田</th>
<th>服</th>
<th>系</th>
<th>统</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>W2</td>
<td>W3</td>
<td>W4</td>
<td>W5</td>
<td>W6</td>
</tr>
</tbody>
</table>

Term: T1 = (W1) T2 = (W1, W2) T3 = (W2) T4 = (W2, W3) T5 = (W3) T6 = (W4)

(2) Multi-Linear Term-Phrasing

<table>
<thead>
<tr>
<th>木</th>
<th>上</th>
<th>田</th>
<th>服</th>
<th>系</th>
<th>统</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>W2</td>
<td>W3</td>
<td>W4</td>
<td>W5</td>
<td>W6</td>
</tr>
</tbody>
</table>

Term: T1 = (W1) T2 = (W1, W2) T3 = (W1, W2, W3) T4 = (W2) T5 = (W2, W3) T6 = (W4)

Fig. 6. Single-linear versus multi-linear term- phrasing for "Desktop (W1-W2) Publishing (W3-W4) Systems (W5-W6)".

45450 terms were extracted from the bibliographic database, with 1-word, 2-word, 3-word, and 4-word phrases.

C. Legality Checking

Term- phrasing often caused an undesired shadow effect for Chinese information. In order to clean up some of the noise, a built-in dictionary was used. As discussed early, in modern Chinese technical literature, a single Chinese word plays the role of a character in English. In English, permutations of words in a meaningful term-phrase are often still meaningful. For example, "Desktop Publishing Systems" can produce meaningful adjacent term-phrases like "Desktop Publishing" or "Publishing Systems". But this is not true for Chinese phrases. For example, the equivalent "桌上出版系统" actually consists of three meaningful Chinese terms: "桌上" (Desktop), "出版" (Publishing), and "系统" (Systems). But the term-phrasing process may create invalid permutations like "桌上出版", "出版系统" and so on. These invalid term-phrases were undesired shadows of the valid term-phrases.

Illegal phrases should be removed from the source entry. As described in [C1], [S0], [I7], [S5], [S3], a built-in dictionary could be employed to solve the word-division problem. Our system included a built-in dictionary to check the legality of the extracted terms. Table II presents a comparison of the statistics generated by using of the built-in dictionary and the result of legality checking for the sample bibliographic data. By applying only 3303 terms in the dictionary we were able to screen out 22139 illegal terms. The terms in the built-in dictionary were created manually based on the dictionary described in [C1] and some common computer science terms. Online, general-purpose dictionary and other domain-specific dictionaries (e.g., in business, computer science, engineering, medicine, etc.) are often available from vendors and publishers or can be generated through OCR scanning of some existing hard-copy sources. In this experiment we created our dictionary manually because of our limited domain and the prototype nature of our experiment. However, in larger-scale applications, existing dictionaries could be incorporated.

Since the modern Chinese language no longer treats a word as a basic semantic unit, single Chinese words were not considered as valid descriptors and only legal 2-word, 3-word, and 4-word phrases were used in concept space generation. English terms were not analyzed in this phase; single-word English terms were included in the concept space generation process.

D. Filtering

After removing illegal Chinese terms, legal but non-semantic-bearing (general) Chinese terms were then filtered by a stopterm list, a process similar to stopwording in English. The dictionary used in legality checking was mainly based on [C1] and included many general entries that were not useful for indexing purposes. Table III summarizes the results of stopterm filtering. A total of 1581 stop terms were included and they helped filter 3791 terms in our sample database. (We could also have created a smaller, but more precise dictionary for both legality checking and filtering purposes.) Fig. 5 displays some sample stop terms.

By applying the legality checking function and a stopterm file of 1581 distinct terms, the collection of terms created from automatic indexing was reduced from 45450 to 5126 (4078 Chinese terms and 1048 English terms). Almost 89% of noisy terms derived from the multi-linear term- phrasing phase
TABLE III
STATISTICS OF TERM FILTERING

<table>
<thead>
<tr>
<th>Term</th>
<th>2-word terms</th>
<th>3-word terms</th>
<th>4-word terms</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stopterm List</td>
<td>793</td>
<td>514</td>
<td>274</td>
<td>1581</td>
</tr>
<tr>
<td>Non Stop Terms in Database</td>
<td>2493</td>
<td>544</td>
<td>1041</td>
<td>4078</td>
</tr>
<tr>
<td>Stop Terms identified</td>
<td>2399</td>
<td>1058</td>
<td>334</td>
<td>3791</td>
</tr>
</tbody>
</table>

was eliminated. These 5126 terms included only 2126 distinct terms. Each of these represented a unique concept descriptor for the bibliographic database. Fig. 7 shows sample results of the complete automatic indexing procedure.

V. AUTOMATIC CONCEPT CLASSIFICATION

Earlier research on automatic classification has considered a wide range of applications including biological, anthropological, archaeological, and linguistic. In addition to existing statistical techniques [28], [4], many techniques consider class definitions, class-finding, and quasi-classification structures based on various theoretical perspectives [30].

Due to the uncertainty involved in information processing, object-class relationships may change over time. Researchers in information retrieval have been concerned with grouping or clustering documents or descriptors (terms) [16], [19]. Document clustering groups documents of similar nature; while term clustering could be used for automatic thesaurus generation or information abstraction.

In the next two subsections, we focus our discussion on a neural network approach to term classification. The first subsection describes the relevance weighting mechanism in measuring the degree of relevance of pairs of concept descriptors in the concept space. The second subsection describes a neural network model which clustered relevant concept descriptors into concept groups.

A. Relevance Weighting

In order to identify the "relevance" relationship between any of the 2126 distinct concepts identified earlier, each concept descriptor needed to be evaluated further to determine its "importance" in a document and across the whole database. The technique adopted was based on Salton’s Vector Space Model [43] which considered term frequency \( tf_{ij} \) and document frequency \( df_j \). We also used the technique recently in [9].

1) First, a threshold for the document frequency (i.e., number of documents in the database that contain the term \( j \) \( df_j \)) is required to eliminate concept descriptors with low frequency. In our system, this threshold can be decided by the system user. By applying the threshold, a portion of the underlying information may be lost (referred to as information loss). That is, infrequent indexes that appear only in selected source bibliographic entries may no longer be used in the next concept space generation and clustering phase. Despite the potential danger of missing some infrequent but important concept descriptors, the threshold helped remove many noisy terms and significantly improved classification efficiency, an observation confirmed in our earlier study [9]. It should be noted that the document frequency of each term was weighted according to the word count of the specified term. That is, if a term consisted of 3 words, then the original document frequency was multiplied by 3 to get the new document frequency. This weighting scheme helped bring out the often more specific multiple-word terms.

2) A combined term weight \( d_{ij} \) was then computed as follows:

\[
d_{ij} = tf_{ij} \times \log df_j
\]

where \( d_{ij} \) represents the relevance weight of the \( j \)th concept descriptor in document \( i \). This procedure transformed the source document into a representation of a finite set of concept descriptors with relevance weights (between 0 and 1) [36], [31], [22]. This computation is a variant of the SMART indexing function developed by Salton [43], [45]. Because of the characteristics of the small sample Chinese-English bibliographic data (the total number of records in the database was smaller than most public-accessible, large-scale bibliographic databases), we used document frequency (more frequent terms are weighted more heavily) instead of the inverse document frequency (more specific terms are weighted
more heavily) used in large database applications for computing term weights. By using document frequency, we were able to weight concept descriptors that appear in more documents more heavily than descriptors that appear sparsely in the whole database. This simple modification helped the system identify important concepts/terms for representing the documents in the database. In our experience, inverse document frequency was found useful for generating specific indexes for large-scale databases [12], but for smaller applications, we found the document frequency weight to be better for capturing important, consensual concepts [9].

3) A co-importance weight \( d_{ijk} \) was then computed based on the following formula:

\[
d_{ijk} = t_{fijk} \times \log d_{fjk}
\]

where \( d_{ijk} \) represents the relevance weight of descriptors \( j \) and \( k \) in document \( i \), \( t_{fijk} = \min(t_{fij}, t_{fik}) \), and \( d_{fjk} = \min(d_{fj}, d_{fk}) \).

4) An asymmetric co-occurrence function developed by the authors [10] as shown below computed the relevance of any two concept descriptors \( j \) and \( k \) in the concept space (relational weight).

\[
W_{jk} = \frac{1}{n} \sum_{i=1}^{n} \frac{d_{ijk}}{d_{ij}}
\]

\[
W_{kj} = \frac{1}{n} \sum_{i=1}^{n} \frac{d_{ijk}}{d_{ik}}
\]

where \( n \) represents the total number of documents in the database. An empirically determined weight threshold of 0.1 was then adopted to retain strongly-associated pairs of descriptors in the concept space. Fig. 8 shows a portion of such a concept space, where concept pairs are displayed in decreasing order. More relevant concepts showed higher relevance probabilities.

B. Parallel Classification

In order to cluster relevant concepts into concept groups, we adopted a new and interesting connectionist (neural network) approach.

Clustering algorithms [16], [14], [29] have been used extensively in IR over the past two decades [41], [43] because the ability of clustering methods to categorize or classify by assigning items to automatically created groups gives it a natural affinity with the goals of online information processing and management. Clustering methods are usually categorized according to the type of cluster structure they produce. The simple nonhierarchical methods divide the data set into clusters where no overlap is allowed [45]. The more popular hierarchical methods produce a nested data set in which pairs of items or clusters are successively linked until every item in the data set is connected. The most commonly used hierarchical clustering methods are based on either the single-link method which joins, at each step, the most similar pair of objects that are not yet in the same cluster or the complete-link method which uses the least similar pair between each of two clusters to determine the inter-cluster similarity [41]. Despite the usefulness of the hierarchy (also called dendrogram [41]) produced by such methods, these algorithms do not allow incorporation of a priori decisions about the number of desired clusters, cluster size, and criteria for cluster membership and the dendrogram does not provide a meaningful abstract (summary) representation of the data set.

More recently neural network clustering has attracted significant interest from researchers [32], [48]. Neural network clustering offers the ability to determine the size, shape, number, and placement of pattern clusters adaptively while intrinsically operating in parallel [47], [48]. Kohonen’s self-organizing feature maps (SOM) [32] is one of the several neural network clustering algorithms which have demonstrated significant utility in various engineering, scientific, and business applications [33]. In this research we adopted the Hopfield network [27] for information classification. The summation
function used in the algorithm allowed our system to cluster strongly-related terms (concepts) together to form concept groups and a sigmoid transformation function allowed overlapping clusters to be generated during the parallel clustering process. By controlling various thresholds, we were able to generate a small number of meaningful concept groups. We had implemented such an algorithm successfully in previous research which involved clustering ideas generated by meeting participants in an electronic meeting environment [9]. Details about this algorithm are presented below.

A neural network model simulates the self-organizing and adaptive properties of a neurological subsystem in human brain. The neuron nodes and weighted links in a neural network model are active processing agents. The Hopfield neural network [27], [33], which resembles an associative network and exhibits a parallel relaxation property in particular, can transform a partial, noisy distributed pattern into a stable state representation. This important property has been used in various pattern recognition and image restoration applications [33], [21]. In the proposed multilingual classification model, a variant of the Hopfield neural network was developed to cluster highly relevant concept descriptors. The strength of the connection between two neuron nodes revealed the degree of relevance between these two neurons.

The classifying behavior of this network was considered a variant of the competitive learning paradigm of neural networks [13], [25]. Multiple neurons, instead of a single winner neuron, were activated by a specific input neuron. These relevant output neurons, which were all concept descriptors in the concept space, were categorized as a concept group [9]. As shown in Fig. 9, a concept space can be perceived as a single-layer network of relevant (weighted) concepts (nodes). By activating each concept in turn and converging to its strongly-associated neighbors, we could generate overlapping regions (clusters) of concept groups (represented as oval in Fig. 9). The Hopfield network clustering procedure adopted as was follows:

1) **Assigning Connection Weights**
   
   Training of the Hopfield net was performed by the relevance weighting computation phase described earlier. Each concept descriptor was represented by a neuron node and relevance weight was considered as synaptic weight.

2) **Initialization with Unknown Input Pattern**
   
   \[ \mu_i(0) = x_i, \quad 0 \leq i \leq n - 1 \]
   
   \( \mu_i(t) \) is the output of neuron \( i \) at time \( t \) and \( x_i \) which has a value between 0 and 1, indicates a value for neuron \( i \). Initially all neurons were assigned 0 except for the specific node (concept) to be classified (weight of 1). After each iteration, the output computed was then used as the input for the next iteration. This initialization and activation process was repeated \( n \) times (for all \( n \) nodes), each time started with a specific neuron. The activation equation was iterated until the network converged or until it reached 100 iterations.

3) **Activation and Iteration**

   \[ \mu_j(t + 1) = f_s \left( \sum_{i=0}^{n-1} W_{ij} \mu_i(t) \right), \quad 0 \leq j \leq n - 1 \]

   where \( W_{ij} \) is the relational weight defined earlier and \( f_s \) is the continuous sigmoid transformation function [33] as shown below.

   \[ f_s(\text{net}_j) = \frac{1}{1 + \exp \left( -\frac{\text{net}_j - \theta_j}{\theta_0} \right)} \]

   where \( \text{net}_j = \sum_{i=0}^{n-1} t_{ij} \mu_i(t), \theta_j \) served as a threshold output and \( \theta_0 \) was used to alter the shape of the sigmoid function.

4) **Convergence**

   The above process was repeated until there was no change in terms of output in the output layer between two iterations, which was accomplished by checking:

   \[ \sum_{j=0}^{n-1} |\mu_j(t + 1) - \mu_J(t)|^2 \leq \epsilon \]

   where \( \epsilon \) was the maximal allowable difference between two iterations. Once the network converged, the final output represented the set of terms relevant to the starting term. In our system, the following values were used: \( \theta_j = 0.1, \theta_0 = 0.01 \) and \( \epsilon = 1 \), all determined empirically.

The Hopfield net classification process strongly relied upon the associative property of the neural net and the parallel relaxation method for concept activation. A system testing session was performed, of which a detailed description will be provided in the next section.

**VI. SYSTEM IMPLEMENTATION AND EVALUATION**

- **Classification and Abstraction for Information Sources:** In order to obtain a concise categorization of the bibliographic data, an *information loss* (the percentage of documents left unindexed by the concept groups) analysis based on document frequency threshold was performed. The experimental result is shown in Table IV. Thirty-one experiments were carried out for 31 different document frequency thresholds. Numbers of documents and terms left were also computed. This experimental session was conducted on a SPARC 390 workstation.
TABLE IV
CLASSIFICATION AND INFORMATION LOSS ANALYSIS

<table>
<thead>
<tr>
<th>Doc-Freq Threshold</th>
<th># of Documents Indexed</th>
<th># of Terms Left</th>
<th># of Categories Classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>61.56%</td>
<td>47</td>
<td>24</td>
</tr>
<tr>
<td>29</td>
<td>61.75%</td>
<td>48</td>
<td>21</td>
</tr>
<tr>
<td>28</td>
<td>65.70%</td>
<td>51</td>
<td>24</td>
</tr>
<tr>
<td>27</td>
<td>66.16%</td>
<td>54</td>
<td>27</td>
</tr>
<tr>
<td>26</td>
<td>66.96%</td>
<td>56</td>
<td>28</td>
</tr>
<tr>
<td>25</td>
<td>67.82%</td>
<td>59</td>
<td>31</td>
</tr>
<tr>
<td>24</td>
<td>67.82%</td>
<td>59</td>
<td>28</td>
</tr>
<tr>
<td>23</td>
<td>68.88%</td>
<td>68</td>
<td>34</td>
</tr>
<tr>
<td>22</td>
<td>68.88%</td>
<td>68</td>
<td>36</td>
</tr>
<tr>
<td>21</td>
<td>70.91%</td>
<td>73</td>
<td>39</td>
</tr>
<tr>
<td>20</td>
<td>71.39%</td>
<td>72</td>
<td>44</td>
</tr>
<tr>
<td>19</td>
<td>73.22%</td>
<td>90</td>
<td>51</td>
</tr>
<tr>
<td>18</td>
<td>74.08%</td>
<td>91</td>
<td>51</td>
</tr>
<tr>
<td>17</td>
<td>75.72%</td>
<td>102</td>
<td>56</td>
</tr>
<tr>
<td>16</td>
<td>75.72%</td>
<td>102</td>
<td>55</td>
</tr>
<tr>
<td>15</td>
<td>78.33%</td>
<td>120</td>
<td>53</td>
</tr>
<tr>
<td>14</td>
<td>78.52%</td>
<td>123</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>79.00%</td>
<td>131</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>79.58%</td>
<td>133</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>82.27%</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>82.95%</td>
<td>181</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>83.82%</td>
<td>195</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>84.39%</td>
<td>214</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>87.28%</td>
<td>320</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>87.48%</td>
<td>321</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>90.27%</td>
<td>400</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>90.75%</td>
<td>414</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>95.47%</td>
<td>1044</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>97.21%</td>
<td>1284</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>98.94%</td>
<td>1872</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>100.00%</td>
<td>2126</td>
<td></td>
</tr>
</tbody>
</table>

Since the major focus of our system evaluation was to assess its capability to extract and cluster relevant concept descriptors for multilingual information, the traditional performance measures of recall and precision [43] were either inappropriate or impractical. To obtain a smaller number of concept groups (between 20 and 40 concept groups in our implementation) to represent the key contents in the sample bibliographic database, an information loss ratio of 30% was adopted. That is, we retained only frequently occurring terms in the database (which represent about 70% of the documents) and used them in Hopfield network clustering, an extensive computational process. This selection was different from the 10% loss reported in [9] because the number of electronic meeting comments was only about 300 and it was desirable to retain as many comments as possible. In our sample bibliographic database, a total of 1052 documents were included.

As shown in Table IV, 34 concept groups represented 68.88% of documents in the sample database when the document frequency threshold was set to 23. For different domains and applications, the above information loss analysis may need to be performed in order to determine the appropriate document frequency threshold, a process which is pretty straight-forward. The 34 concept groups represented the majority of the subjects discussed in the documents of the bibliographic database. However, the complete indexes generated earlier can still be used for keyword-based searching. The information loss analysis only intended to "abstract" and represent the key topics in the database, i.e., it was used for classification purposes.

Classifying and representing information sources succinctly could contribute significantly to the success of the recent internet resource discovery services, as tens of thousands of information sources (e.g., bibliographic databases, bulletin boards, etc.) become available for search on Internet [39]. Creating a directory to summarize the services provided by individual information sources has been a consistent challenge for researchers. We believe our approach provides a viable alternative to the prevailing keyword indexing approach to classification. However, a more detailed comparison of the performance of our approach versus other existing techniques for large-scale information sources still needs to be performed.

Fig. 10 shows a graphical display of the concept classification analysis. It can be seen that when the document frequency threshold was increased, there was a sharp decline in the number of terms left, but the number of document indexed and the number of categories declined gradually. We believe this suggests that terms with high document frequency tend to represent a majority of the concepts in a large database. However, this postulation requires a systematic testing across different domains and for different databases. The result of automatic concept classification is presented in Fig. 11.

As shown in Fig. 11, relevant concept descriptors were grouped by the Hopfield algorithm to form a concept category. The more general concept categories were ranked higher in the classification results. Because the ETEN magazine is mainly a forum for discussing computer science research and applications in Taiwan, the categories classified by the Hopfield neural network revealed many important topics discussed in the database such as: 1: Chinese/ETEN Chinese/ETEN Chinese systems; 2: software; 3: hard disk management/design/management/program design...; 4: assembler/assembly; 5: network/local area network; 6: input/input method/Chinese input; 7: compiler; 8: database; 9: graphics; 10: C language/C/mouse, 11: memory chip/memory, 12: development; 13: printer/printing/color/laser printing, 14: Chinese computer, 15: editing; 16: application program, 17: character/Chinese character; 18: interface/graphics, and so on. A topic may be discussed extensively in several documents and it also is likely that some documents may involve multiple topics. As is made evident in Fig. 11, some concept groups included 2-9 terms which were strongly related and both Chinese and English descriptors were captured and clustered. All concept descriptors also appeared relevant and precise.

• Concept-Based Information Retrieval: In addition to automatic classification, the concept space generated as a result of co-occurrence analysis, as appears in Fig. 8, also provided an added functionality for supporting concept-based information retrieval. Using the proposed tech-
niques, our system generated a total of 2126 distinct terms for the sample database and on average each term has about 7 neighboring (related) terms. The terms and their weighted relationships can be perceived as a system-generated thesaurus, which represents the important concepts and their cross-reference structure in the underlying database (we refer to this as a concept space).

In the event of an unsuccessful search using his/her own terms, a searcher can consult a system-generated concept space to identify other semantically relevant search terms, a thesaurus consultation process prominent in expert searchers' IR behaviors [6]. Often complementing the conventional keyword search capability (based on full-text or automatic indexing), thesaurus search has also been incorporated into many prevailing full-text retrieval software packages, e.g., BRS/SEARCH², BASIS/Plus³, and Topic⁴. We believe that with the extensive indexing capabilities provided by such full-text retrieval software we can adopt the automatic concept space generation capability of our proposed approach in the full-text retrieval environment. The thesaurus browsing and concept-augmentation features available in full-text retrieval software have enormous potential for use in a system equipped with an automatically-generated, domain-specific thesaurus. It should be noted that no

² Vended by BRS Software Products, McLean, VA.
³ Vended by Information Dimensions Inc., Dublin, OH.
⁴ Vended by Verity, Inc., Mountain View, CA.
existing full-text retrieval software exhibits concept space generation capability and the commercial packages have only limited capability in processing multilingual information.

VII. CONCLUSIONS AND FUTURE EXTENSIONS

Database management systems (DBMS) have provided a feasible solution to indexing and classifying structured, numeric information. However, for unstructured, textual information, development of an automatic and “intelligent” information system is important. This research proposed a systematic blueprint of a multilingual classification model to help automatically index and classify unstructured Chinese-English information.

Special features of this information classification model include a Chinese multi-linear term-phrasing technique, an integrated Chinese-English automatic indexing framework, and a neural network based concept classification component. For automatic multilingual indexing, the multi-linear term-phrasing method adopted to extract term-phrases for a concept space of Chinese and English descriptors. In automatic concept classification, a variant of the Hopfield neural network and its parallel relaxation algorithm was developed to categorize concept descriptors. We believe this multilingual methodology can also contribute to research related to other multilingual databases involving languages such as Spanish-English or Japanese-English. The specific directions for our future research include:

- Developing a more complete built-in dictionary, and stopterm list for legibility checking and filtering of the term-phrases extracted from multi-linear term-phrasing.
- Implementing the multilingual information classification model on a parallel machine. Both legibility checking and spreading activation of Hopfield neural network could be implemented more efficiently in a parallel mode.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their comments and suggestions.

REFERENCES


CHINESE REFERENCES

(1) 楊慶元等人， “現代語詞庫應用詞頻分析方法”， 訊號出版， 北京， 1979. 6

(2) 阮明基， “循序漸進： 說要對ICP出版本點的看法”， (明誠工業學報)， 38 期 77, pp. 89–90

(3) 郭信榮， “中文與英文筆記型數學系統比較研究”， (中華民國學會)， 48 期 77, pp. 89–90

(4) 羅鵬， “中文實驗系統研究與分析”， (中華民國學會)， 40 期 76, pp. 101–105

(5) 宋玉， “自動化語義解析系統的研究及設計”， (中華民國學會)， 45 期 78, pp. 27–31

(6) 鄭志強， “中文語義的處理”， (中華民國學會)， 19 期 78, pp. 101–111

(7) 陳俊軒， “電腦語義語義系統的研究及設計”， (中華民國學會)， 45 期 78, pp. 27–31

(8) 陳俊軒， “電算化語義系統”， (中華民國學會)， 19 期 78, pp. 88–91

(9) 楊志華， “中文語義系統比較”， (中華民國學會)， 45 期 78, pp. 58–59

(10) 董明勳， “中文語義系統的基礎”， (中華民國學會)， 19 期 78, pp. 28–29

(11) 丁之元， “語義資料庫： 一個中文語義系統的實例”， (中華民國學會)， 19 期 78, pp. 28–29

(12) 杨惠文， “中文語義庫的發展路向”， (中華民國學會)， 19 期 78, pp. 128–154

(13) 楊浩華， “中文後處理”， (中華民國學會)， 33 期 71, pp. 238–243

(14) 楊浩華， “中文語義系統”， (中華民國學會)， 33 期 71, pp. 196–201

(15) 楊浩華， “中文後處理”， (中華民國學會)， 33 期 71, pp. 238–243

Chung-hsin Lin received the M.S. degree in management information systems from the University of Arizona, Tucson, in 1992. His research interests include databases and Chinese information indexing and retrieval.

Hsinchun Chen received the Ph.D. degree in information systems from the Leonard N. Stern School of Business, New York University, in 1989. He is an Associate Professor of Management Information Systems at the Karl E. Hail Graduate School of Business, University of Arizona, Tucson. His research interests include TCSW, human-computer interactions, text-based information management and retrieval, multilingual information retrieval, internet resource discovery, knowledge acquisition and knowledge discovery, machine learning, and neural network modeling and classification.