

# 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]<sup>1</sup>. For exam-

Manuscript received March 5, 1993; revised April 24, 1994, and October 7, 1994. This project was supported in part by a grant awarded by the *International Program Development Fund*, University of Arizona, 1992–1993, and a Research Initiation Award grant awarded by the Division of Information, Robotics, and Intelligent Systems, National Science Foundation (IRI-9211418), 1992–1994.

The authors are with the Department of Management Information Systems, Karl Eller Graduate School of Management, University of Arizona, Tucson, AZ 85721.

Publisher Item Identifier S 1083-4419(96)00409-8.

<sup>1</sup> English references are listed in arabic numbers. For Chinese references, the reference number is prefixed with 'C', e.g., [C1].

ple, 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 grammatic 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 associated 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 problems 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 effective 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 information has become significantly greater, especially in large-scale databases. This situation is particularly evident for textual databases, which are widely used in traditional library science 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 knowledge, system knowledge, and classification scheme knowledge are able to maneuver and explore in these textual databases [7].

In conventional information retrieval environments, keywords 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 vocabulary 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], presents an alternative approach for handling information retrieval 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 example, 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 allows 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], meeting 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, animations, 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, stopwording, 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 underlying 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].

( $m > 0$ )	ATIONAL	→	ATE	relational	→	relate
( $m > 0$ )	TIONAL	→	TION	conditional	→	condition
				rational	→	rational
( $m > 0$ )	ENCY	→	ENCE	valency	→	valence
( $m > 0$ )	ANCY	→	ANCE	hesitancy	→	hesitance
( $m > 0$ )	IZER	→	IZE	digitizer	→	digitize
( $m > 0$ )	ABLY	→	ABLE	conformably	→	conformable
( $m > 0$ )	ALLY	→	AL	radically	→	radical
( $m > 0$ )	ENTLY	→	ENT	differently	→	different
( $m > 0$ )	ELY	→	E	vilely	→	vile
( $m > 0$ )	OUSLY	→	OUS	analogously	→	analogous
( $m > 0$ )	IZATION	→	IZE	vietnamization	→	vietnamize
( $m > 0$ )	ATION	→	ATE	predication	→	predicate
( $m > 0$ )	ATOR	→	ATE	operator	→	operate
( $m > 0$ )	ALISM	→	AL	feudalism	→	feudal
( $m > 0$ )	IVENESS	→	IVE	decisiveness	→	decisive
( $m > 0$ )	FULNESS	→	FUL	hopefulness	→	hopeful
( $m > 0$ )	OUSNESS	→	OUS	callousness	→	callous
( $m > 0$ )	ALITY	→	AL	formality	→	formal
( $m > 0$ )	IVITY	→	IVE	sensitivity	→	sensitive
( $m > 0$ )	BILITY	→	BLE	sensibility	→	sensible

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.

隨身的伙伴—膝上型電腦，李華企劃〔倚天〕18 民78.08 pp.40  
 麻雀雖小五臟全，編輯部〔倚天〕18 民78.08 pp.42  
 膝上型電腦與中文搭配，劉春暉〔倚天〕18 民78.08 pp.50  
 海闊天空任遨遊，李華〔倚天〕18 民78.08 pp.52  
 如何選購可攜式電腦，李森華〔倚天〕18 民78.08 pp.55  
 市場縮地，藍麗霞〔倚天〕18 民78.08 pp.58  
 沒有電線的區域網路—清爽潔淨的辦公環境，邱惠玲〔倚天〕18 民78.08 pp.66  
 IBM 推出新版OS/2 VERSION 1.2，蔡志豪編譯〔倚天〕18 民78.08 pp.68  
 簡介 IBM PC BIOS，胡定一譯，〔倚天〕18 民78.08 pp.71  
 CPU 的新紀元80486，68040，萬正祥編譯〔倚天〕18 民78.08 pp.121  
 機械製圖的新領域—電腦輔助設計，藍麗霞企劃〔倚天〕18 民78.08 pp.124  
 談CAD/CAM 專用網路系統，盧崇仁，〔倚天〕18 民78.08 pp.126

Fig. 2. Sample Chinese-English bibliographic 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 to 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 ETEN 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 ETEN 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

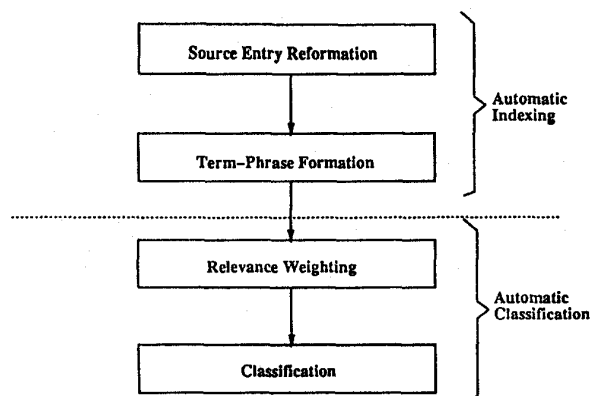


Fig. 3. English classification model.

the unstructured part (i.e., title) of each bibliographic entry. A total of 1052 document entries were stored in the sample bibliographic database.

### 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-

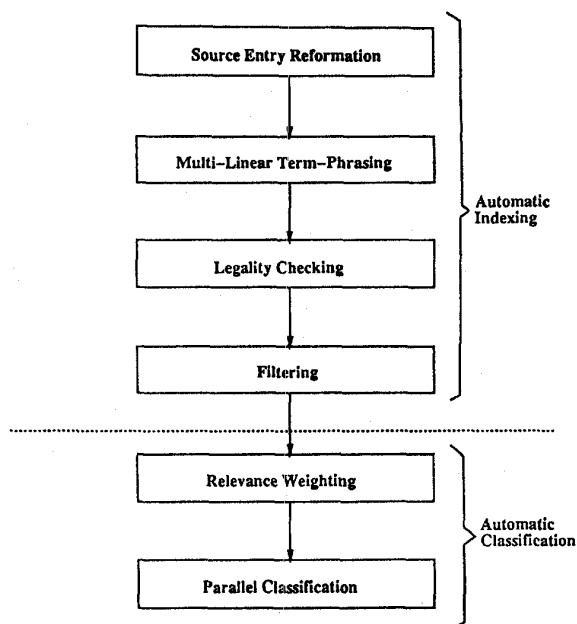


Fig. 4. Multilingual classification model.

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

所有 (All)  
 業者 (Practitioner)  
 努力 (Effort)  
 使用 (Use)  
 困擾 (Trouble)  
 大家 (Everyone)  
 選購 (Purchase)  
 要素 (Factor)  
 市場 (Market)  
 再生 (Re-birth)  
 新力量 (New Force)  
 為什麼 (Why)  
 大家談 (Everyone Talks)  
 爭霸戰 (Fighting War)  
 新世界 (New World)  
 新計畫 (New Plan)  
 追求卓越 (Pursuit of Excellence)  
 系統概論 (System Introduction)  
 正本清源 (Final Clarification)  
 衆所矚目 (Everyone Watches)  
 蓄势待發 (Ready to Go)  
 特殊意義 (Special Meaning)  
 設計要件 (Design Factor)  
 展翅高飛 (Soaring Above)  
 世界一家 (World Village)  
 資金調度 (Capital Movement)

Fig. 5. Chinese stopterm list.

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 *meaning unit filter* knowledge base in [1], was used to perform this 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

TABLE I  
RESULTS OF MULTI-LINEAR TERM-PHRASING

1-word terms	2-word terms	3-word terms	4-word terms	Total
15442	12580	9947	7481	45450

our efforts to resolve the automatic Chinese-English indexing problem. Examples and analyses are also presented.

##### A. Source Entry Reformation

Source entry reformation 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.

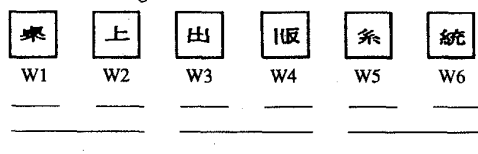
##### B. Multi-Linear Term-Phrasing

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.

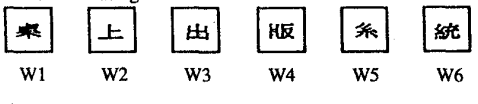
Fig. 6 shows the different possible permutations of adjacent words which can be generated for the same text segment using the two approaches. Table I presents a summary of the results of multi-linear term-phrasing for the bibliographic database;

(1) Single-Linear Term-Phrasing



- Term:  
 T1 = (W1)    T7 = (W1, W2)  
 T2 = (W2)    T8 = (W3, W4)  
 T3 = (W3)    T9 = (W5, W6)  
 T4 = (W4)    T10 = (W1, W2, W3)  
 T5 = (W5)    T11 = (W4, W5, W6)  
 T6 = (W6)

(2) Multi-Linear Term-Phrasing



- Term:  
 T1 = (W1)    T7 = (W1, W2)    T12 = (W1, W2, W3)  
 T2 = (W2)    T8 = (W2, W3)    T13 = (W2, W3, W4)  
 T3 = (W3)    T9 = (W3, W4)    T14 = (W3, W4, W5)  
 T4 = (W4)    T10 = (W4, W5)    T15 = (W4, W5, W6)  
 T5 = (W5)    T11 = (W5, W6)    T16 = (W1, W2, W3, W4)  
 T6 = (W6)    T17 = (W2, W3, W4, W5)  
 T18 = (W3, W4, W5, W6)

Fig. 6. Single-linear versus multi-linear term-phrasing for "Desktop 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 "桌上出版系統" actually consists of three meaningful Chinese terms: "桌上" (Desktop), "出版" (Publishing), and "系統" (Systems). But the term-phrasing process may create invalid permutations like "桌上出", "版系統" and so on. These invalid term-phrases were undesired *shadows* of the valid terms-phrases.

Illegal phrases should be removed from the source entry. As described in [C1], [50], [17], [35], [53], a built-in dictionary could be employed to solve the word-division problem. Our

TABLE II  
 STATISTICS OF LEGALITY CHECKING

Term	2-word terms	3-word terms	4-word terms	Total
Dictionary	1633	810	895	3338
Legal Terms in Sample Database	4892	1602	1375	7869
Illegal Terms in Sample Database	7688	8345	6106	22139

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 3338 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



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:

$$d_{ijk} = tf_{ijk} \times \log df_jk$$

where  $d_{ijk}$  represents the relevance weight of descriptors  $j$  and  $k$  in document  $i$ ,  $tf_{ijk} = \min(tf_{ij}, tf_{ik})$ , and  $df_jk = \min(df_j, df_k)$ .

- 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).

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

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

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 clus-

---

倚天 : 中文 = 0.68918919 (Eten : Chinese)	
倚天 : 倚天中文 = 0.62162162 (Eten : Eten Chinese)	
倚天 : 中文系統 = 0.26440264 (Eten : Chinese Systems)	
倚天 : 設計 = 0.15409487 (Eten : Design)	
倚天 : 硬碟管理 = 0.13220132 (Eten : Hard Disk Management)	
倚天 : 管理 = 0.12737733 (Eten : Management)	
倚天 : 硬碟 = 0.11948735 (Eten : Hard Disk)	
中文 : 倚天 = 0.29038546 (Chinese : Eten)	
中文 : 倚天中文 = 0.27720851 (Chinese : Eten Chinese)	
中文 : 中文系統 = 0.17267676 (Chinese : Chinese Systems)	
中文電腦 : 中文 = 1.00000000 (Chinese Computer : Chinese)	
中文電腦 : 發展 = 0.25000000 (Chinese Computer : Development)	
滑鼠 : C = 0.26666667 (Mouse : C)	
滑鼠 : C 語言 = 0.26666667 (Mouse : C Language)	
網路 : 區域網路 = 0.13531425 (Network : Local Area Network)	
管理 : 硬碟管理 = 0.59615385 (Management : Hard Disk Management)	
管理 : 硬碟 = 0.55922699 (Management : Hard Disk)	
管理 : 管理模組 = 0.23334567 (Management : Management Module)	
管理 : 設計 = 0.23076923 (Management : Design)	
管理 : 模組 = 0.21086728 (Management : Module)	
管理 : 倚天 = 0.19230769 (Management : Eten)	
管理 : 中文 = 0.19230769 (Management : Chinese)	
管理 : 中文系統 = 0.19230769 (Management : Chinese Systems)	
管理 : 倚天中文 = 0.19230769 (Management : Eten Chinese)	
管理 : 管理系統 = 0.13354269 (Management : Management Systems)	
管理 : 模組設計 = 0.11480303 (Management : Modular Design)	
管理系統 : 管理 = 1.00000000 (Management Systems : Management)	
管理系統 : 資料庫 = 0.33333333 (Management Systems : Database)	
管理系統 : 網路 = 0.11111111 (Management Systems : Network)	
管理系統 : 區域網路 = 0.10745911 (Management Systems : Local Area Network)	
中文系統 : 中文 = 1.00000000 (Chinese Systems : Chinese)	
中文系統 : 倚天 = 0.64516129 (Chinese Systems : Eten)	
中文系統 : 管理 = 0.31080977 (Chinese Systems : Management)	

---

Fig. 8. Partial result of co-relevance weighting.

ters 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 ovals in Fig. 9). The Hopfield network clustering procedure adopted was as 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

$$\mu_i(0) = x_i, 0 \leq i \leq n - 1$$

$\mu_i(t)$  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.

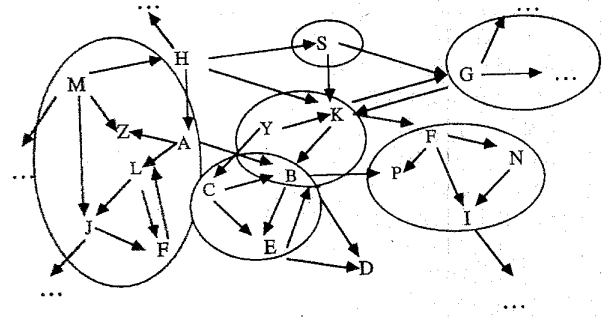


Fig. 9. A conceptual diagram for Hopfield net clustering.

#### 3) Activation and Iteration

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

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

Doc-Freq Threshold	# of Documents Indexed	# of Terms Left	# of Categories Classified
30	61.56%	47	24
29	61.75%	48	21
28	65.70%	51	24
27	66.18%	54	27
26	66.96%	56	28
25	67.82%	59	31
24	67.82%	59	28
23	68.88%	68	34
22	68.88%	68	36
21	70.91%	73	39
20	71.39%	72	44
19	73.22%	90	51
18	74.08%	91	51
17	75.72%	102	56
16	75.72%	102	55
15	78.32%	120	53
14	78.52%	123	-
13	79.00%	131	-
12	79.58%	133	-
11	82.27%	180	-
10	82.95%	181	-
9	83.82%	195	-
8	84.39%	214	-
7	87.28%	320	-
6	87.48%	321	-
5	90.27%	409	-
4	90.75%	414	-
3	95.47%	1044	-
2	97.21%	1284	-
1	98.94%	1872	-
0	100.00%	2126	-

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 sciences 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/graphs, 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-

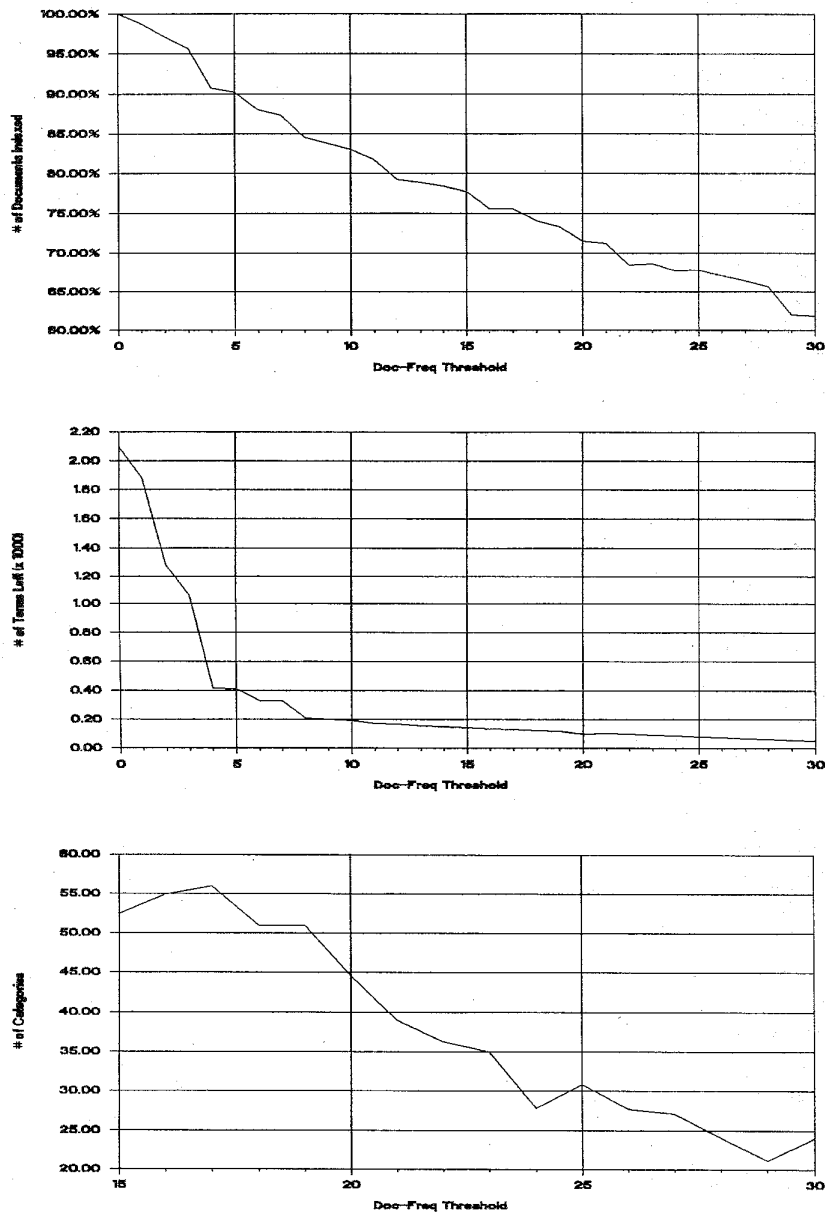


Fig. 10. Concept classification analysis.

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/SEARCH<sup>2</sup>, BASIS/Plus<sup>3</sup>, and Topic<sup>4</sup>. 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

<sup>2</sup>Vended by BRS Software Products, McLean, VA.

<sup>3</sup>Vended by Information Dimensions Inc., Dublin, OH.

<sup>4</sup>Vended by Verity, Inc., Mountain View, CA.

---

1 :	中文/倚天中文/倚天/中文系統 (Chinese/Eten Chinese/Eten/Chinese Systems)
2 :	軟體 (Software)
3 :	硬盤管理/設計/管理/程式設計/硬碟/管理模組/模組/管理系統/模組設計 (Hard Disk Management/Design/Management/Program Design/Hard Disk/ Management Module/Module/Management Systems/Modular Design)
4 :	組合語言/組合 (Assembly/Assembler)
5 :	網路/區域網路 (Network/Local Area Network)
6 :	輸入/輸入法/中文輸入 (Input/Input Method/Chinese Input)
7 :	編譯 (Compiling)
8 :	資料庫 (Database)
9 :	繪圖 (Graphics)
10 :	C 語言/C/滑鼠 (C Language/C/Mouse)
11 :	記憶體/記憶 (Memory Chip/Memory)
12 :	發展 (Development)
13 :	印表機/印表/彩色/雷射印表 (Printer/Printing/Color/Laser Printing)
14 :	中文電腦 (Chinese Computer)
15 :	編輯 (Editing)
16 :	應用程式 (Application Programs)
17 :	文字/中文字 (Character/Chinese Character)
18 :	介面/圖形 (Interface/Graphs)
19 :	病毒/電腦病毒 (Virus/Computer Virus)
20 :	技巧 (Technique)
21 :	視窗/視窗系統 (Windows/Windows Systems)
22 :	多媒體/媒體 (Multimedia/Media)
23 :	LOTU (LOTU)
24 :	CLIPPER (CLIPPER)
25 :	工具 (Tool)
26 :	WINDOW (WINDOW)
27 :	BASIC (BASIC)
28 :	DO/程式語言 (DO/Programming Language)
29 :	TURBO PASCAL/TURBO (TURBO PASCAL/TURBO)
30 :	行列輸入 (Line Input)
31 :	掃描器 (Scanner)
32 :	電腦輔助 (Computer-aided)
33 :	新翰藝 (New Han-E)
34 :	專家系統 (Expert Systems)

---

Fig. 11. Classification/clustering results.

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 was 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 legality checking and filtering of the term-phrases extracted from multi-linear term-phrasing.
- Implementing the multilingual information classification model on a parallel machine. Both legality 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

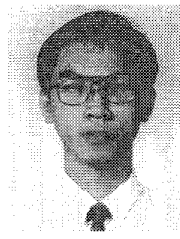
- [1] A. L. Baker, J. M. Bieman, and W.F. Tooley, "Interlingua: A multi-language business information center," *Proc. 21st Annual Hawaii Int. Conf. System Sciences*, vol. 4, pp. 79-86, 1988.
- [2] D. C. Blair, "Indeterminacy in the subject access to documents," *Inform. Proc. Mgmt.*, vol. 22, no. 2, pp. 229-241, 1986.
- [3] E. Carmel, S. Crawford, and H. Chen, "Browsing in hypertext: A cognitive study," *IEEE Trans. Syst. Man Cyber*, vol. 22, no. 5, pp. 865-884, Sept./Oct. 1992.
- [4] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman, "Autoclass: A Bayesian classification system," in *Proc. Fifth Int. Conf. Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1988.
- [5] H. Chen, "Collaborative systems: Solving the vocabulary problem," *IEEE Computer*, vol. 27, no. 5, pp. 58-66, Special Issue on Computer-Supported Cooperative Work (CSCW), May 1994.
- [6] H. Chen and V. Dhar, "Reducing indeterminism in consultation: A cognitive model of user/librarian interaction," in *Proc. 6th National Conf. Artificial Intelligence (AAAI-87)*, Seattle, WA, July 13-17, 1987, pp. 285-289.
- [7] ———, "User misconceptions of online information retrieval systems," *Int. J. Man-Machine Studies*, vol. 32, no. 6, pp. 673-692, June 1990.
- [8] ———, "Cognitive process as a basis for intelligent retrieval systems design," *Inform. Proc. Mgmt.*, vol. 27, no. 5, pp. 405-432, 1991.
- [9] H. Chen, P. Hsu, R. Orwig, L. Hoopes, and J. F. Nunamaker, "Automatic concept classification of text from electronic meetings," *Commun. ACM*, vol. 37, no. 10, Oct. 1994.
- [10] H. Chen and K. J. Lynch, "Automatic construction of networks of concepts characterizing document databases," *IEEE Trans. Syst. Man Cyber*, vol. 22, no. 5, pp. 885-902, Sept./Oct. 1992.
- [11] H. Chen, K. J. Lynch, K. Basu, and T. Ng, "Generating, integrating, and activating thesauri for concept-based document retrieval," *IEEE Expert, Special Series on Artificial Intelligence in Text-Based Information Systems*, vol. 8, no. 2, pp. 25-34, Apr. 1993.
- [12] H. Chen, B. Schatz, T. Yim, and D. Fye, "Automatic thesaurus generation for an electronic community system," in *J. Amer. Soc. Inform. Sci.*, vol. 46, no. 3, pp. 175-193, Apr. 1995.
- [13] T. E. Doszkocs, J. Reggia, and X. Lin, "Connectionist models and information retrieval," *Ann. Rev. Inform. Sci. Technol. (ARIST)*, vol. 25, pp. 209-260, 1990.
- [14] R. Dubes and A. K. Jain, "Clustering methodologies in exploratory data analysis," *Advances in Computers*, M. C. Yovits, Ed., vol. 19, pp. 113-228, 1980.
- [15] O. Ekeberg, "Robust dictionary lookup using associative networks," *Int. J. Man-Machine Studies*, vol. 28, pp. 29-43, 1988.
- [16] B. Everitt, *Cluster Analysis*, second edition. London, England: Heinemann, 1980.
- [17] C. K. Fan and W. H. Tsai, "Automatic word identification in Chinese sentences by the relaxation technique," *Comp. Proc. Chinese and Oriental Languages*, vol. 4, no. 1, pp. 33-56, 1988.
- [18] Y. Feng and K. Z. Wang, "Chinese question-answer experimental system based on the sense coherence among the SNEs," *Second Intl. Conf. Computers and Applications*, 1987, pp. 927-933.
- [19] D. Fisher, "Knowledge acquisition via incremental conceptual clustering," *Machine Learning*, vol. 2, pp. 139-172, 1987.
- [20] C. L. Foss, "Tools for reading and browsing hypertext," *Inform. Proc. Mgmt.*, vol. 25, no. 4, pp. 407-418, 1989.
- [21] L. Fu, *Neural Networks in Computer Intelligence*. New York, NY: McGraw-Hill, 1994.
- [22] N. Fuhr and C. Buckley, "A probabilistic learning approach for document indexing," *ACM Trans. Inform. Syst.*, vol. 9, no. 3, pp. 223-248, July 1991.

- [23] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary problem in human-system communication," *Commun. ACM*, vol. 30, no. 11, pp. 964-971, Nov. 1987.
- [24] L. S. Gay and W. B. Croft, "Interpreting nominal compounds for information retrieval," *Inform. Proc. Mgmt.*, vol. 26, no. 1, pp. 21-38, 1990.
- [25] S. Grossberg, "Competitive learning: From interactive activation to adaptive resonance," *Cogn. Sci.*, vol. 11, pp. 23-63, 1987.
- [26] D. Gusfield, G. M. Landau, and B. Schieber, "Efficient algorithm for the all pairs suffix prefix problem," *Inform. Proc. Lett.*, vol. 41, no. 4, pp. 181-185, 1992.
- [27] J. J. Hopfield, "Neural network and physical systems with collective computational abilities," *Proc. National Academy of Science, USA*, vol. 78, no. 8, pp. 2554-2558, 1982.
- [28] L. Hunter and D. J. States, "Bayesian classification of protein structure," *IEEE Expert*, Aug. 1992, pp. 67-74.
- [29] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1988.
- [30] K. S. Jones, "Some thoughts on classification for retrieval," *J. Documentation*, vol. 26, no. 2, pp. 89-101, June 1970.
- [31] ———, "Experiments in relevance weighting of search terms," *Inform. Proc. Mgmt.*, vol. 15, pp. 133-144, 1979.
- [32] T. Kohonen, *Self-Organization and Associative Memory*, 3rd Ed. Berlin Heidelberg: Springer-Verlag, 1989.
- [33] R. P. Lippmann, "An introduction to computing with neural networks," *IEEE ASSP Mag.*, vol. 4, no. 2, pp. 4-22, Apr. 1987.
- [34] L. B. Lovins, "Development of a stemming algorithm," *Mechan. Translat. Computat. Linguistics*, vol. 11, pp. 22-31, 1968.
- [35] J. H. Lu, "Approach to converting phrases in pinyin to phrases in [Chinese characters]," *Proc. 1988 IEEE Int. Conference Syst. Man Cyber.*, vol. 1, pp. 389-391, 1988.
- [36] Z. Mazur, "Properties of a model of information retrieval system based on thesaurus with weights," *Inform. Proc. Mgmt.*, vol. 15, pp. 145-154, 1979.
- [37] R. S. Michalski and R. E. Stepp, "Learning from observation: Conceptual clustering," in *Machine Learning, An Artificial Intelligence Approach*, R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, Eds. Palo Alto CA: Tioga, 1983, pp. 331-363.
- [38] J. Nielsen, *Hypertext and Hypermedia*. New York, NY: Academic Press, 1990.
- [39] K. Obraczka, P. B. Danzig, and S. Li, "Internet resource discovery services," *IEEE Computer*, vol. 26, no. 9, pp. 8-24, Sept. 1993.
- [40] M. E. Porter, *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. New York: Free Press, 1980.
- [41] E. Rasmussen, "Clustering algorithms," in *Information Retrieval: Data Structures and Algorithms*, W. B. Frakes and R. Baeza-Yates, Eds. Englewood Cliffs, NJ: Prentice Hall, 1992.
- [42] G. Salton, "Generation and search of clustered files," *ACM Trans. Database Systems*, vol. 3, no. 4, pp. 321-346, Dec. 1978.
- [43] ———, *Automatic Text Processing*. Reading, MA: Addison-Wesley, 1989.
- [44] G. Salton, J. Allan, and C. Buckley, "Automatic structuring and retrieval of large text files," *Commun. ACM*, vol. 37, no. 2, pp. 97-108, Feb. 1994.
- [45] G. Salton and M. E. Lesk, "Information analysis and dictionary construction," in *The Smart Retrieval System—Experiments in Automatic Document Processing*, G. Salton, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1971, pp. 115-142.
- [46] C. B. Schwind, "Semantic trees for natural language representation," *Inform. Proc. Mgmt.*, vol. 19, no. 4, pp. 223-235, 1983.
- [47] P. K. Simpson, *Artificial Neural Systems: Foundations, Paradigms, Applications, and Implementations*. New York, NY: McGraw-Hill Book Company, 1990.
- [48] ———, "Fuzzy min-max neural networks—Part 2: Clustering," *IEEE Trans. Fuzzy Systems*, vol. 1, no. 1, pp. 32-45, Feb. 1993.
- [49] K. Sparck Jones, *Automatic Keyword Classification for Information Retrieval*. London: Butterworths, 1971.
- [50] R. Sproat and C. L. Shih, "A statistical method for finding word boundaries in [Chinese text]," in *Proc. 1989 Int. Conf. Computer Processing of Chinese and Oriental Languages*, Changsha, China, 1990.
- [51] R. E. Stepp, "Concepts in conceptual clustering," in *Proc. 10th Int. Joint Conf. Artificial Intelligence (IJCAI-87)*, Milan, Italy, Aug. 23-28, 1987, pp. 211-213.
- [52] S. S. Tseng, C. C. Yang, and C. C. Hsieh, "An experimental model of Chinese textual database," *J. Chinese Inst. Engineers*, vol. 13, no. 6, pp. 607-622, 1990.
- [53] X. D. Wang and B. Q. Dai, "Chinese speech understanding system," *9th Int. Conf. Pattern Recognition*, 1988, pp. 14-17.
- [54] C. W. Yang, Z. Lai, and Y. Zhang, "Application of Chinese natural language understanding to robot control," *Proc. 1988 IEEE Int. Conf. Syst. Man Cyber.*, vol. 1, pp. 385-388, 1988.

## CHINESE REFERENCES

- (1) 廖慶琪等人, "現代漢語常用詞詞類詞典", 宇航出版社, 北京, 民79.6
- (2) 游萬來, 蔡登傳, "應用kwic法編製中文設計期刊索引的研究", (明志工專學報) 20 民77.05 pp.9-24
- (3) 顧敏, "中英文資訊儲存暨檢索系統比較研究", (中國圖書館學會會報) 43 民77.12 pp.97-106
- (4) 顧敏, "立法資訊系統文獻詞彙索引方法", (中國圖書館學會會報) 40 民76.06 pp.101-106
- (5) 宋玉, "自動化資訊檢索系統中索引的設計", (中國圖書館學會會報) 45 民78.12 pp.27-31
- (6) 謝清俊, "中文資訊的處理 -1- -2- -3-", (科學月刊) 18:2 民76.02 pp.149-152, 18:3 民76.03 pp.230-232, 18:5 民76.05 pp.383-386
- (7) 陳悅表, "我國法律全文檢索系統的設計與開發", (資訊與電腦) 90 民77.01 pp.102-111
- (8) 陳悅表, "電腦化資訊檢索系統", (資訊與電腦) 79 民76.04 pp.69-71
- (9) 沙燕琪, "全文檢索系統比較", (資訊傳真) 94 民77.05 pp.58-59
- (10) 謝清俊, "全文檢索的方法", (科學月刊) 19:4 民77.04 pp.262-267
- (11) 丁之侃, "史籍資料庫:一個中文全文處理系統的實例", (科學月刊) 19:4 民77.04 pp.268-272
- (12) 徐惠文, "全文資料庫的發展與現況", (科學月刊) 19:4 民77.04 pp.128-154
- (13) 簡立峰, 卜小蝶, "文件資料庫與文字資料檢索", (倚天雜誌) 51 民81.05 pp.238-243
- (14) 簡立峰, 卜小蝶, "全文檢索", (倚天雜誌) 52 民81.06 pp.196-201
- (15) 簡立峰, 卜小蝶, "中文全文檢索", (倚天雜誌) 53 民81.07 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 Eller Graduate School of Business, University of Arizona, Tucson. His research interests include CSCW, 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.

He received an NSF Research Initiation Award in 1992 and the Hawaii International Conference on System Sciences (HICSS) Best Paper Award in 1994. He was awarded a Digital Library Initiative grant by NSF/NASA/ARPA (1994-1998) recently. Dr. Chen has published more than 20 articles in publications such as *Communications of the ACM*, *IEEE COMPUTER*, *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS*, *IEEE EXPERT*, *Journal of the American Society for Information Science*, *Information Processing and Management*, *International Journal of Man-Machine Studies*, and *Advances in Computers*.