8/10 2004 _ 15pages _ 2021.11.03

- self-organizing map SOM-Algorithm
- phrase extraction through frequency reduction
- map visualization is also interesting

ELSEVIER

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Decision Support Systems 39 (2005) 583–597

Decision Support Systems

www.elsevier.com/locate/dsw

Newsmap: a knowledge map for online news

Thian-Huat Ong[a,*], Hsinchun Chen[b], Wai-ki Sung[b], Bin Zhu[c]

[a] *Department of Management Information Science, College of Business Administration, California State University, Sacramento, USA*
[b] *Department of Mangement Information System, Eller College of Management, University of Arizona, USA*
[c] *Information System Department, School of Management, Boston University, USA*

Available online 18 May 2004

Abstract

Information technology has made possible the capture and accessing of a large number of data and knowledge bases, which in turn has brought about the problem of *information overload*. Text mining to turn textual information into knowledge has become a very active research area, but much of the research remains restricted to the English language. Due to the differences in linguistic characteristics and methods of natural language processing, many existing text analysis approaches have yet to be shown to be useful for the Chinese language. This research focuses on the automatic generation of a hierarchical knowledge map NewsMap, based on online Chinese news, particularly the finance and health sections. Whether in print or online, news still represents one important knowledge source that people produce and consume on a daily basis. The hierarchical knowledge map can be used as a tool for browsing business intelligence and medical knowledge hidden in news articles. In order to assess the quality of the map, an empirical study was conducted which shows that the categories of the hierarchical knowledge map generated by NewsMap are better than those generated by regular news readers, both in terms of recall and precision, on the sub-level categories but not on the top-level categories. NewsMap employs an improved interface combining a 1D alphabetical hierarchical list and a 2D Self-Organizing Map (SOM) island display. Another empirical study compared the two visualization displays and found that users' performances can be improved by taking advantage of the visual cues of the 2D SOM display.

Keywords: Knowledge management; Knowledge map; Online news; Neural networks; Experimental research

1. Introduction

The information technology think tank Gartner Group defines Knowledge Management (KM) as: "a discipline that promotes an integrated approach to identifying, capturing, retrieving, sharing and evaluating an enterprise's information assets" [16]. Current information technology ably enables people to capture and access large amounts of information in structured and semi-structured data and knowledge bases, causing there to be more information available than humans can process, a phenomenon commonly referred to as "information overload" [3,7]. To alleviate information overload, current Knowledge Management researchers are applying newer artificial intelligence and visualization techniques to extract and visualize knowledge from the mass of information.

[*] Corresponding author. Tel.: +1-916-278-6023.
E-mail addresses: ongt@csus.edu (T.-H. Ong), hchen@eller.arizona.edu (H. Chen), wai-ki@eller.arizona.edu (W. Sung), bzhu@bu.edu (B. Zhu).

0167-9236/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2004.03.008

Information users usually employ one of two strategies to locate information of interest: searching and browsing. First, the users search when they already have in mind a topic or some keywords. Full-text information retrieval systems [35] and Internet search engines such as Google and Alta Vista are typical examples of systems that deal extensively with searching problems. Second, users browse when they do not have a specific thing they want to look for, whether it be an unfamiliar area in which they are interested and want to explore or something that has aroused curiosity. The research reported here primarily focuses on the browsing aspect of information seeking, because providing hierarchical subject categories (akin to Yahoo! directory) or knowledge maps (knowledge structures represented by a map metaphor) has been shown to be an effective way to support browsing behaviors [5,8,20,25,26].

The explosion of information has made manual efforts to create subject categories or knowledge maps an overwhelming task. Furthermore, normal information users often are not capable of navigating through a large space of information. An automatic approach to creating subject categories or knowledge maps therefore is necessary to provide relief for both the information creators and information users. In order to create such knowledge structures from diverse information and knowledge sources, Gartner Group has suggested a bottom-up approach that includes data extraction, linguistic analysis, dictionary/thesaurus creation, semantic networks, clustering/categorization, and concept yellow pages [16]. Our research follows this bottom-up approach by extracting relevant phrases from a news collection using a statistical phrase extractor, hierarchical categorizing, and visualizing the knowledge maps.

The challenges of this research are to create high-quality hierarchical knowledge maps and to create effective visualizations for those knowledge maps. This research adopts an automatic approach to generating a hierarchical knowledge map for knowledge sources, in particular Chinese news sources. Through knowledge representation using Updateable PAT-Tree Phrase Extraction, clustering analysis using Self-Organizing Map, and knowledge map visualization, our research aims to alleviate the problem of information overload.

2. Automatic knowledge map for Chinese news: literature review

A map is a drawing that reveals physical and/or abstract relationships for places or objects of interest. Throughout human history, people have been creating physical maps, such as cave paintings, atlases, and more recently satellite scans and three-dimensional computer visualization. Examples of abstract maps include Concept Map for learning objects [29] and Mind Map for improving memorization [4]. A knowledge map is a knowledge representation that reveals the underlying relationships of the knowledge sources, using a map metaphor for spatial display. For example, a knowledge map for news articles could highlight the current major news topics and their relationships by using blocks to represent key concepts and possibly using lines to represent relationships.

Although users' interactions with an information technology system could be a valuable source of information, this research assumes that there exists content from which a map can be derived, either directly or readily obtainable from the Internet. The following subsections review existing approaches to generating knowledge maps, characteristics of news, and why adaptation to Chinese content is important.

2.1. Knowledge map systems

2.1.1. Subject hierarchy

A subject hierarchy or directory is an alphabetical list of topics organized into groups and subgroups. Although lacking the visualization of a map, it is a simple yet effective way to categorize a large volume of information, especially when coupled with a search function, such as telephone yellow pages and the Library of Congress classification system. Representative examples on the Internet include Yahoo! (http://www.yahoo.com) and the Open Directory Project (http://www.dmoz.org). However, limitations of an alphabetic display have been recognized in Refs. [15,17,27]. The main disadvantage of list display is that the interface itself becomes increasingly difficult for a user to navigate as the hierarchy grows larger. Furthermore, the relationships between topics on the same level cannot be expressed in a linear list.

2.1.2. Manual knowledge maps

Concept Map [29] and Mind Map [4] are drawings in which blocks represent concepts or things and connecting lines represent relationships. These maps are used to help better organize, understand, and memorize knowledge. However, the creation process is very personal and demanding of the map creator's cognitive skills. To lighten the manual process, Shneiderman [36] has identified eight activities that support creativity and has proposed the GenEx (Generator of Excellence) framework to help designers create powerful tools that enable users to be more creative more of the time. However, the manual approach is not scalable to the processing of large amounts of information, because a manual knowledge map is not only limited in scope and timeliness, but it is also slow and cumbersome. In order to create a knowledge map that closely matches a mental model, an intelligent, automatic categorization algorithm must be employed.

2.1.3. Automatic knowledge maps

Dodge and Kitchin [13] did a comprehensive survey of maps of cyberspace (cybermaps) that have been generated since the inception of the Internet and categorized them into three categories: mapping infrastructure and traffic, mapping the Web, and mapping conversation and community. Similarly, automatic knowledge maps can be categorized into three categories based on their knowledge characteristics.

2.1.3.1. A. Numerical. Visualization of numbers was among the earliest map applications. When the numbers have physical correspondence, the maps are easily understood. For example, the Internet's statistics on inbound and outbound traffic and the domain host's density can be readily layered on top of the physical infrastructure maps to reveal the growth and demographics of the Internet over time (http://www.telegeography.com). In contrast, visualizing the financial landscape does not have any immediate physical meaning, so visualization algorithms need to be employed to provide meaning to the position and size on the map. For example, SmartMoney.com uses Ordered Treemap to place about 600 companies such that similar companies are placed close to each other and companies with higher market capitalization are given a bigger size [37].

2.1.3.2. B. Textual. Mapping textual knowledge sources is more difficult than mapping numerical knowledge sources because text has limited spatial meaning but strong abstract or conceptual relationships. One example of mapping textual knowledge is Cartia's NewsMap [12], which applies sophisticated and proprietary lexical algorithms to analyze and, in some senses "understand", the content of text documents and the relations between them to distill the key topics and form the topic island map. Another common approach is to use neural networks, which are a class of machine learning algorithms consistent with human mental models [22]. In particular, the Kohonen Self-Organizing Map (SOM) is an unsupervised neural networks algorithm that does not require a predetermined pattern in order to organize a large amount of information into categories such that similar categories are placed close to each other, much as the brain would do. Lin et al. [21] were some of the first to adopt single-layer SOM to information retrieval, by clustering important concepts in a small database of 140 abstracts related to artificial intelligence. He found that SOM is able to create a semantic map that captures the relationships between concepts. However, the concepts are subject descriptors manually indexed by librarians. One of the earlier automatic hierarchical knowledge map generation systems, reported in Ref. [6], was for categorizing a portion of the Internet web pages contained in the Yahoo! Entertainment sub-category (about 110,000 web pages). A subsequent empirical study in Ref. [8] showed that the system could successfully categorize a large amount of information into a meaningful hierarchy of subject maps. A prototype system showed that both fisheye and fractal views could increase the effectiveness of the SOM visualization [42]. In addition, WEBSOM was created to generate a large one-level SOM map based on 1 million newsgroup documents [18]; the map provided a zoomable but not a hierarchical structure.

2.1.3.3. C. Social. The third kind of knowledge maps visualizes human relationships in a community. Rich knowledge sources may include numerical, textual, and relationship links. Social visualization research represents human behavior graphically. For example, systems such as Loom [14] and PeopleGarden [40] provide a graphical representation of who starts a discussion, who talks with whom, how long a person stays, and how lively a discussion is. Chat Circles [14] aims to facilitate synchronous conversation in an online chat room, where each participant is assigned a circle and the size of the circle indicates the freshness of a posting. Users can only "hear" from or "talk" to others in their vicinity. They can also move their circles around to find an appropriate subgroup to talk with. However, our research focused on textual knowledge maps because our goal was to generate knowledge maps from a large textual knowlegde collection.

2.2. *Internet news portals*

News is an important knowledge source for creating knowledge maps, since most people still rely on reading the news to gather new information and useful knowledge. News represents important knowledge about human activities on a daily basis, ranging from political and economic topics to health and science. For business people, the finance section of a newspaper is a must to keep abreast of the current business environment and activities. For general readers who are not health-care professionals, the health section of a newspaper provides adequate coverage of health issues and knowledge relevant to living a healthy lifestyle. Compared with normal web pages, news is also an excellent text mining testbed because it is usually written by a select group of reporters with strong editorial support to ensure high quality.

News Portals, such as Yahoo! News and Excite News, act as an intermediary to deliver the news created by news services such as Associated Press and Reuters with value-added services and customization [23]. Some news portals, such as CNN and MSNBC, also house their own writing staffs. However, current technology like Google focuses mainly on sophisticated searching capabilities and personalization based on self-selected categories or on collaborative filtering [19] and relevant links to the news that the readers are reading.

Currently, newspapers only broadly divide the news into different sections, which should be sufficient if the readers quickly browse through the headlines to look for interesting news on a particular day. However, the division into sections is insufficient when the readers are trying to understand a particular topic over a period of time.

One important value-added service that a news portal can provide is helping readers understand news content. For instance, news editors routinely categorize interesting news into special topics to attract readers. In the wake of the Enron investigation in the US, they might put together several news articles related to auditing practices from the past to the present, to give readers a coherent story to explain and predict how those practices will affect accounting conduct in the future.

Such special topics are valuable knowledge sources, but creating them is laborious and at times costly. They demand a lot of an editor's cognitive skills and time. Most importantly, the topics are very limited in scope because they specifically target a particular segment of their audience and are of little use for the rest of the users who are interested in other topics.

In sum, the manual approach is costly and unable to keep up with the enormous and growing volume of news. Therefore, an automatic approach can enhance this valuable service by providing a hierarchical knowledge map to allow readers to browse different evolving topics in a timely manner.

2.3. *Chinese content*

According to an AC Nielsen survey released on April 22, 2002, the United States has the highest number of Internet users at home: 166 million people. China is second with around 56.6 million households [31]. However, the number represents only 5.5% of the Chinese population. In addition, an earlier study by China Internet Network Information Center (CNNIC) released on February 20, 2001 showed that 76% of Chinese Internet users access Chinese-language information [30]. Therefore, the need for Chinese information sources is huge and growing fast.

Information Retrieval research has a long tradition in English, whereas IR research in Chinese is relatively new. The foundation of Information Retrieval is *indexing*, the process of representing a document with a vector of terms [35]. Multi-word phrases are preferred to single words, because phrases capture a richer linguistic representation of a document [1]. A Chinese sentence is made up of a consecutive sequence of Chinese characters, so the indexing task becomes extracting the longest meaningful sequence of characters. Unlike English, which has already developed many linguistically based approaches, the Chinese indexing problem is still an active research area, due to the lack of linguistic clues in the Chinese language. Since linguistically based Chinese phrase extractors are not yet successful and a dictionary approach is limited to coverage of new vocabulary, a statistical approach is often adopted for Chinese [9,32,41]. Based on our previous research, we selected a variation of the Updateable PAT-Tree Phrase Extraction approach to extract phrases for indexing purpose. This technique is able to extract long, precise Chinese phrases.

This research aims to apply newer neural-network-based categorization techniques to Chinese knowledge sources. We are optimistic that the automatic approach will work well for Chinese content, because the document representation used by the categorization technique is generic, and our indexing technique is able to provide precise phrases that are representative of the documents.

3. Newsmap: facilitating knowledge browsing over Chinese news

This section explains in detail the analysis algorithms, testbed, and visualization interface for the system. Fig. 1 shows the high-level process for automatically generating hierarchical knowledge maps. The key analysis algorithms are statistically based Chinese Indexing and neural-network-based SOM Categorization. We used Chinese news as our testbed and visualized the results of hierarchical knowledge maps by a combination of Internet browser and Java applet.

3.1. Analysis algorithms

3.1.1. Chinese phrase extractor

Statistically based phrase extraction has roots in collocations, which are defined as arbitrary and recurrent word combinations [10,11,38]. However, longer phrases constructed from bigrams occur more frequently, and the computation model is not scalable. For a domain specific corpus, identifying the rigid noun phrases, which are a type of collocations, is more important, because of specialized vocabulary and usage. Based on efficient data structure and searching algorithms, the PAT-Tree Phrase Extraction Approach [9,32] is an iterative process of identifying significant lexical patterns by examining the statistical frequencies and co-occurrences of sequences of Chinese characters. *Mutual information* is a metric that measures how frequently a pattern occurs in the corpus, relative to its sub-patterns:

$$\mathrm{MI}_c = \frac{f_c}{f_{\mathrm{left}_c} + f_{\mathrm{right}_c} - f_c}$$

Fig. 2 shows the Chinese phrase “artificial intelligence” (人工智慧), which is a pattern of interest. The left and right sub-patterns are partial words that are not meaningful in Chinese. Therefore, they are less likely to occur on their own but are most likely to co-occur with meaningful pattern c. In this case, all three frequencies are equal, so MI_c is high and close to 1, which means pattern c is likely to be a good phrase. On the other hand, if MI_c is low and close to 0, the pattern c is not likely to form a phrase.

The basic algorithmic requirement for this approach to work is that it must be able to calculate the frequencies quickly. That means it needs both an efficient data structure, such as semi-infinite string representation, and an efficient search algorithm, such as PAT-Tree, PAT array, and suffix array, that makes it feasible to analyze a large training corpus [2,24].

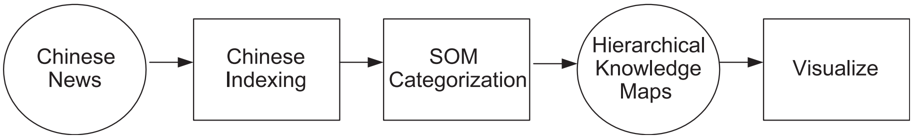


Fig. 1. High-level process for automatic hierarchical knowledge map generation.

c =	人工智慧	“artificial intelligence”
left =	人工智	(partial word, no meaning)
right =	工智慧	(partial word, no meaning)

Fig. 2. A potential pattern for a Chinese phrase and its sub-patterns.

The algorithm first looks for the longest available character sequences. For each length, extract all the possible phrases of a particular length, and repeat this process until there are no more phrases of this length to be extracted. Then the algorithm moves to the next smaller length. The criterion for a pattern to be extracted is to pass the thresholds for predetermined frequency and mutual information value.

After the phrase has been extracted, all its sub-patterns, whether valid or invalid, may also be extracted, which could potentially increase the errors in phrase extraction. Therefore, our extension of updateable data structure [32] supports online updates to decrease the frequency of the extracted phrase pattern. This ensures that the invalid sub-patterns will not survive because frequency reduction of the longer pattern will also decrease the frequency of those invalid sub-patterns. However, the valid sub-patterns may still survive as long as they exist independently and pass the mutual information threshold.

This approach is language-independent in nature because it only cares about the frequency of the co-occurring good phrases. With the appropriate choice of analysis units, such as character for Chinese and word for English or other western languages, one can in theory extend this approach to phrase extraction in many other languages. In our previous research [32], we compared with a PAT-Tree approach [9] and found that our updateable approach improve recall from 0.19 to 0.43 and precision from 0.52 to 0.70.

3.1.2. SOM categorization

In order to generate a hierarchical knowledge map, the system needs to be able to categorize all the indexing terms and group the broader categories on a higher level while putting the narrower subcategories under one of the broader category. In addition, the placement of the categories must be semantically related. Our previous research has concluded that a variant of the Kohonen self-organizing feature map (SOM) is a suitable candidate. Its ability to label the clusters created and its two-dimensional output makes SOM appropriate for visualizing clusters. A scaleable multi-layered graphical SOM approach to Internet categorization developed in our previous research can be used to create an intuitive, graphical display of the important concepts contained in textual information [6,33]. In addition, SOM was found to perform at least as well as other clustering algorithms in clustering text documents [34]. Below we describe the steps of the multi-layered SOM algorithm:

1. Initialize input nodes, output nodes, and connection weights. After removing stopwords, we use the top 1000 frequent Chinese terms extracted from our news testbed as the input vector, because other less-frequent terms are less likely to be selected as main topics. We then create a 20-by-20 two-dimensional map of 400 output nodes to support placement of important topics. Next, we initialize all the connection weights to small random numbers.
2. Present all news articles in order. We represent each news article by a vector of 1000 indexing terms and present them to the system.
3. Compute distances to all nodes. We compute distance d_j between the input and each output node j.
4. Select winning node j^* and update weights to node j^* and neighbors. We select winning node j^* as that output node with minimum d_j. Then we update weights for node j^* and its neighbors to reduce their distances (between input nodes and output nodes).
5. Label regions in map. After the network is trained through repeated presentation of all inputs, we assign a label to each output node by choosing the one corresponding to the indexing term with the largest weight (winning term). Neighboring nodes that contain the same labels are merged to form a region. Similarly, we submit each document as input to the trained network again and assign it to the winning node on the map. The resulting knowledge map thus represents regions of important categories (the more important the concept, the larger the region) and news articles are assigned to their corresponding regions.
6. Apply the above steps recursively for large regions. For each map region that contains more than 100

news articles, we conduct a recursive procedure of generating another self-organizing map until each region contains no more than 100 news articles.

3.2. Testbed: a Chinese news collection

The testbed news collection was provided by the one of the biggest Taiwanese news companies, which publishes seven Chinese newspapers in Taiwan and around the world, both in print and online. The articles were provided in a 600 MB Microsoft Access database and dated from September 1999 to April 2000 for a total of 331,156 news articles. Currently, the articles are assigned into a main section and seven subsections each day. The main section consists of the newspaper's front page and news not assigned to any of the seven subsections. As news articles accumulate over time, it becomes difficult for any user to browse within any subsection, as evidenced by the large number of news articles in Table 1, which incorporates a collection of less than 1 year.

The news primarily focused on Taiwanese local interests; regional and international news is included if sufficient interest is expressed by readers. For example, regional and international politics, finances, and major events are covered regularly. In addition, many readers follow the news of U.S. National Basketball Association (NBA) games in the Sports section and the activities of Hollywood actors and actresses are reported in the Entertainment subsection.

3.3. Knowledge map visualization

The NewsMap visualization interface (Fig. 3) includes both a 1D alphabetical expandable hierarchical list and a 2D SOM island display, because list-oriented users and spatially oriented users may perform better in one than in the other.

Table 1
Distribution of news articles in different sections

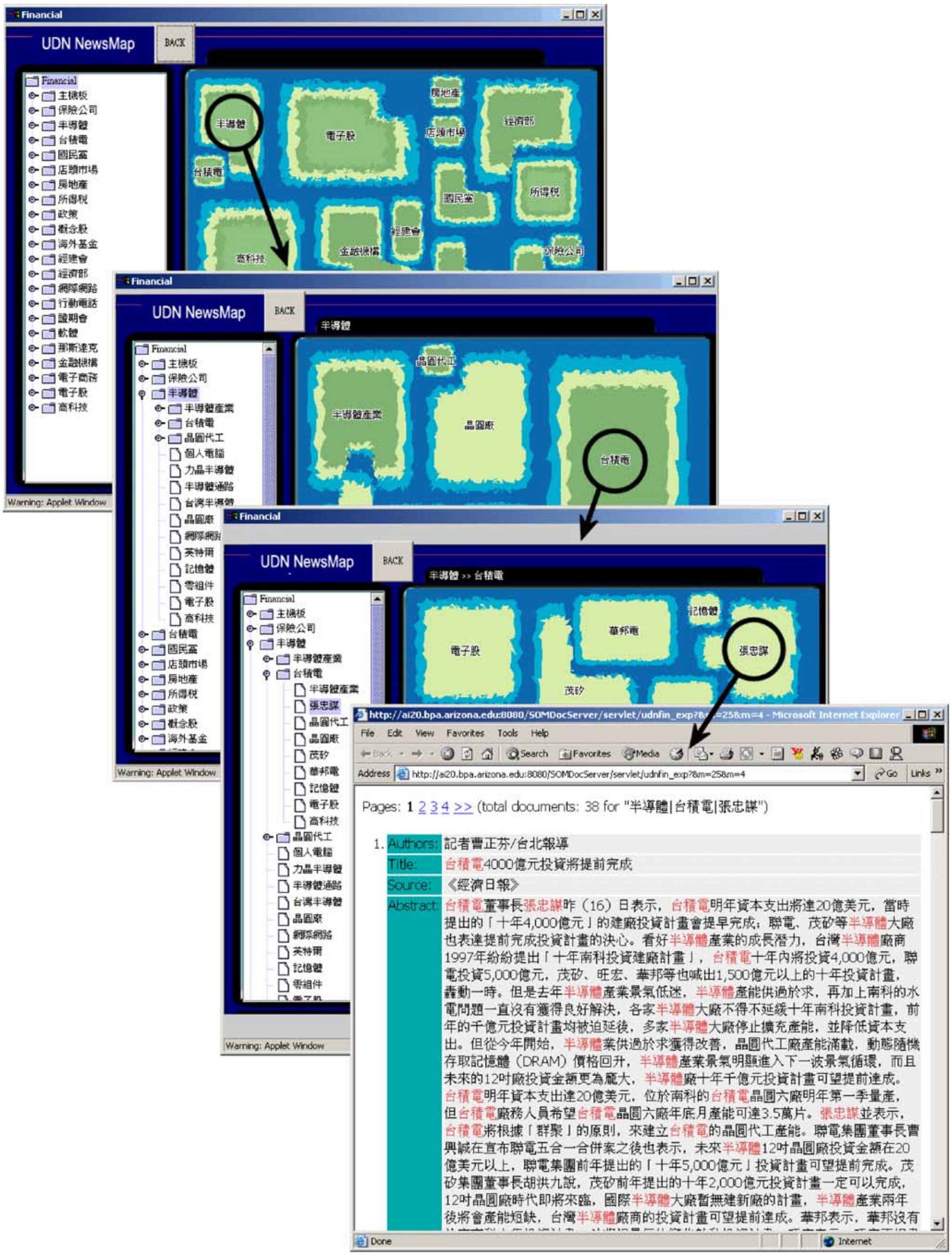
Section	Articles
Main	216,217
(1) Health	6343
(2) Finance	63,386
(3) Sports	18,050
(4) Society	4340
(5) Lifestyle	7956
(6) Entertainment	10,151
(7) Travel	4713
Total	331,156

The advantage of the 2D SOM display is that the spatial proximity between categories corresponds with their semantic proximity. Furthermore, the 2D SOM display is improved by introducing an ocean-and-island metaphor that is more intuitive and conveys more visual clues. Each category appears as an island in the ocean, with a surrounding light blue color to represent shallow water and different layers of green shades to indicate the number of depth in the levels of subcategories. Another visual clue is that the size of the island gives an estimate of the number of news articles contained in the category.

When the user first enters the system, the top-level knowledge map always shows up, which is the map at the top of Fig. 3. The circled category of "semiconductor" (半導體) has three layers of green shades, which indicates that the user can click on the category and go down two more levels. As the user clicks on this category, both the 1D and 2D displays are updated. In the 1D display, the hierarchical list is expanded and the selected category is highlighted. The 2D display is replaced by the newly selected sub-level map for "semiconductor" (半導體). The traversal path is shown just above the 2D display to remind the user of his/her current position. The interface also provides a back button, so the user can get back to the previous screen easily. All the categories at the lowest level are represented by a light green shade, which will bring up the news articles when the user double-clicks on it.

Fig. 3 shows a series of views of the Financial NewsMap, which reacts to the user's clicking on selected categories. On each map, similar categories are placed close to each other. For example, on the top level, "semiconductor" (半導體), "Taiwan Semiconductor Manufacturing" (台積電), "electronic stocks" (電子股), and "high-tech" (高科技) are placed close to each other, because they describe the high-tech industry in Taiwan. On the other hand, "Ministry of Economic Affairs" (經濟部), "income tax" (所得稅), "real estate properties" (房地產), "the Kuomintang" (國民黨), and "Council for Economic Planning and Development" (經建會) are all representative of government business involvement and are placed at the upper right hand corner.

As the user zooms into the second level for "semiconductor" (半導體), other meaningful subcategories show up, such as "Intel" (英特爾), "computer



http://ai20.bpa.arizona.edu:8080/SOMDocServer/servlet/udnfin_exp?&m=25&m=4 - Microsoft Internet Explorer

Address http://ai20.bpa.arizona.edu:8080/SOMDocServer/servlet/udnfin_exp?&m=25&m=4

Pages: 1 2 3 4 >> (total documents: 38 for "半導體|台積電|張忠謀")

1. Authors: 記者曹正芬/台北報導

Title: 台積電4000億元投資將提前完成

Source: 《經濟日報》

Abstract: 台積電董事長張忠謀昨（16）日表示，台積電明年資本支出將達20億美元，當時提出的「十年4,000億元」的建廠投資計畫會提早完成；聯電、茂矽等半導體大廠也表達提前完成投資計畫的決心。看好半導體產業的成長潛力，台灣半導體廠商1997年紛紛提出「十年南科投資建廠計畫」，台積電十年內將投資4,000億元，聯電投資5,000億元，茂矽、旺宏、華邦等也喊出1,500億元以上的十年投資計畫，轟動一時。但是去年半導體產業景氣低迷，半導體產能供過於求，再加上南科的水電問題一直沒有獲得良好解決，各家半導體大廠不得不延緩十年南科投資計畫，前年的千億元投資計畫均被迫延後，多家半導體大廠停止擴充產能，並降低資本支出。但從今年開始，半導體業供過於求獲得改善，晶圓代工廠產能滿載，動態隨機存取記憶體（DRAM）價格回升，半導體產業景氣明顯進入下一波景氣循環，而且未來的12吋廠投資金額更為龐大，半導體廠十年千億元投資計畫可望提前達成。台積電明年資本支出達20億美元，位於南科的台積電晶圓六廠明年第一季量產，但台積電廠務人員希望台積電晶圓六廠年底月產能可達3.5萬片。張忠謀並表示，台積電將根據「群聚」的原則，來建立台積電的晶圓代工產能。聯電集團董事長曹興誠在宣布聯電五合一合併案之後也表示，未來半導體12吋晶圓廠投資金額在20億美元以上，聯電集團前年提出的「十年5,000億元」投資計畫可望提前完成。茂矽集團董事長胡洪九說，茂矽前年提出的十年2,000億元投資計畫一定可以完成，12吋晶圓廠時代即將來臨，國際半導體大廠暫無建新廠的計畫，半導體產業兩年後將會產能短缺，台灣半導體廠商的投資計畫可望提前達成。華邦表示，華邦沒有

Fig. 3. The knowledge map visualization for NewsMap. The example shows the user clicks through different sub-levels to the news articles.

memory" (記憶體), "Personal Computer" (個人電腦), and "the Internet" (網際網路). Next, the user can zoom into the third level for "Taiwan Semiconductor Manufacturing" (台積電), where he/she can see its competitors "Mosel Vitelic" (茂矽) and "Winbond" (華邦電). When the user double-clicks on its CEO "Morris Chang" (張忠謀), it brings up all the news articles with all the labels of selected categories being highlighted in red.

4. Evaluating quality of knowledge map

In order to determine the quality of the knowledge map generated by our approach, we conducted an experiment that compared the system-generated list of categories with those generated by the Taiwanese news readers. Our proposed research question is: Are the categories in the hierarchical knowledge map created by the NewsMap system comparable to categories created by readers?

4.1. Experiment design and procedure

Based on our past research in SOM, we believed that SOM would be able to perform comparably with an average human subject for a categorization task. The reason for this is that when SOM creates a knowledge map, it is more comprehensive by using important indexing terms generated from actual news content. On the other hand, although an average reader has more sophisticated mental processing capability, he/she may generate a map that is less comprehensive because he/she is limited by his/her ability to recall important related topics. Therefore, we hypothesize that NewsMap would produce better topic recall and precision than human readers from actual news articles:

- H1a: NewsMap has better recall at the top level.
- H1b: NewsMap has better recall at the sub-level.
- H2a: NewsMap has better precision at the top level.
- H2b: NewsMap has better precision at the sub-level.

The experiment measured recall and precision of topics generated by our system and human readers. Recall is a measure of thoroughness or the ratio of correct selection to the answer set.

$$\text{subject recall} = \frac{\text{number of categories selected by both subjects and expert}}{\text{number of categories selected by expert}}$$

$$\text{subject recall} = \frac{\text{number of categories selected by both system and expert}}{\text{number of categories selected by expert}}$$

On the other hand, precision is a measure of accuracy or the ratio of correct selection to the selection set.

$$\text{subject precision} = \frac{\text{number of categories selected by both subjects and expert}}{\text{number of categories generated by subject}}$$

$$\text{subject precision} = \frac{\text{number of categories selected by both system and expert}}{\text{number of categories generated by subject}}$$

Since there is no well-established hierarchical knowledge map for news articles that can serve as the benchmark for comparison, the experiment employed two experts to create a standard answer key from which the recall and precision values could be calculated. The experts were a Taiwanese PhD student and an experienced Taiwanese business executive who had lived in Taiwan for more than 20 years, and read Taiwanese news online every day. The procedure used to create the answer key was as follows.

Once all the subjects had completed the experiments, the lists from both the system and the subjects were combined for each separate section on top-level and various sub-levels. In order to make the universe of answers more complete, the domain experts had been asked to come up with their own list of 15 to 25 categories. They then independently created a standard answer key based on their own list and the combined lists. The combined lists were sorted in alphabetically order such that they could not identify whether the categories came from the system or subjects. Some categories were considered equivalent, due to abbreviation or alternate wording. After the experts compiled their own answer keys, they compared and debated with each other to create a standard answer key which was used as the benchmark for comparison.

The experiment tried to limit the scope to only the finance and health subsections, considered to be relevant to most readers. Below is the experiment procedure.

1. To evaluate the top level knowledge map, the task was to ask a subject to write down, from his/her memory, a list of 15 to 25 categories that he/she thought best corresponded to the section of news articles that appeared in newspapers the subject had

read in the past. The range 15 to 25 was chosen because it is comparable to the number of categories generated by the system and it would be sufficient to provide topic coverage.

2. To evaluate the sub-level knowledge map, two categories were randomly chosen from the top-level map, and the task was to ask the subject to list 15 to 25 subcategories that best described each of the two categories.

Each subject was given a total of six tasks, with one top-level task and two sub-level tasks in both the finance and health sections. The experiment subjects performed all the tasks without the help of external resources. The subject was told to take as much time as necessary. The selected sub-level categories for the finance section were “the Internet” (網際網路), “semiconductor” (半導體), “insurance companies” (保險公司), and “Ministry of Economic Affairs” (經濟部), whereas the health section had “diabetes” (糖尿病), “pregnancy” (懷孕), “eye” (眼睛), and “joint” (關節).

Since the news articles originated from Taiwan and contained topics of local interest, the experiment was conducted using 30 Taiwanese students as experiment subjects.

4.2. Results and discussion

In the pilot study, there had been one Chinese student, but we soon found that the vocabulary differences between China and Taiwan were more complex than a difference between simplified and traditional characters. For example, the Chinese name of “Vanguard International Semiconductor” (世界先進) was understood literally as “world advanced” for Chinese people, yet the Taiwanese instantly recognize it as a leading semiconductor company. In fact, these less obvious differences caused some subjects to spend a lot of time figuring out the meaning of the categories. Therefore, the experiment used only Taiwanese students as subjects.

Table 2
Recall measure comparison

Level	System (%)	Human (%)	*N*	Pairwise *p*-value
Top-level	20.9	19.3	60	0.1298
Sub-level	24.3	18.0	120	0.0000*

[*] Statistically significant.

Table 3
Precision measure comparison

Level	System (%)	Human (%)	*N*	Pairwise *p*-value
Top-level	15.6	19.9	60	0.0012*
Sub-level	32.1	21.2	120	0.0000*

[*] Statistically significant.

After the experts had produced the standard answer key, recall and precision were calculated for both system and subjects.

Tables 2 and 3 show that the data supported all the hypotheses with a confidence level of 0.05.

Table 4 shows the categories selected by the experts, the system, and a sample subject for the sub-level category “semiconductor” (半導體). The top half of the table shows the matches between the experts and the system or the subject, and the bottom half simply lists the other non-matching categories. The categories selected by the experts were representative of the semiconductor industry.

4.2.1. Recall

The difference between system recall and human recall was not significant on the top level (rejecting H1a), but was significant on the sub-level (accepting H1b). We believe the reason was that on the top level, the potential pool of candidates was larger so the subjects had more difficulty in recalling the categories from their memory. However, on the sub-level, the subjects had less difficulty because they were focusing on a more specific category.

For example, Table 4 shows seven categories selected by the system matched the experts’ selection. Only five categories selected by the subject matched the experts’ selection. Therefore, the system recall was higher than the subject recall, because the system was able to select from top 1000 indexing terms during the training, while the subject was limited by his/her own knowledge and working memory. Interestingly, two additional categories, “Winbond” (華邦電) and “Morris Chang” (張忠謀, chairman of Taiwan Semiconductor Manufacturing) later appeared in one of the system-generated subcategories “Taiwan Semiconductor Manufacturing Company” (台積電).

4.2.2. Precision

The system precision is significantly lower than human precision on the top level (rejecting H2a), but

Table 4
Comparing the terms selected by expert, system, and a sample subject for the category "semiconductor"

Expert	System	Subject
半導體產業 (semiconductor business)	半導體產業 (semiconductor business)	
半導體通路(semiconductor supplie	半導體通路 (semiconductor supplier)	
台積電 (Taiwan Semiconductor Manufacturing Company)	台積電 (Taiwan Semiconductor Manufacturing Company)	台積電 (Taiwan Semiconductor Manufacturing Company)
晶圓 (wafer), 八吋晶圓 (8 inch wafer), 十二吋晶圓 (12 inch wafer)		晶圓 (wafer)
晶圓代工 (fables foundry)	晶圓代工 (fables foundry)	
晶圓廠 (fabrication foundry)	晶圓廠 (fabrication foundry)	
電子股 (electronic stocks)	電子股 (electronic stocks)	
零組件 (zero component)	零組件 (zero component)	
聯電 (United Microelectronics Corp)		聯電 (United Microelectronics Corp)
員工配股 (stocks option)		股票 (stock)*
新竹科學園區 (Hsinchu Science-based Industrial Park)		新竹 (Hsinchu)*, 科學園區 (Science-based Industrial Park)*
世界先進 (Vanguard)	力晶半導體 (Powerchip)	市場佔有率 (market share)
半導體設計 (semiconductor design)	台灣半導體 (Taiwan Semiconductor)	年產量 (annual production)
張忠謀 (Morris Chang)	英特爾 (Intel)	期業間諜 (enterprise spy)
曹興誠 (Robert Tsao)	個人電腦 (PC)	技術升級 (technical upgrade)
晶片封裝 (chip package)	記憶體 (memory)	電子工廠 (electronic factory)
華邦電 (Winbond)	高科技 (high-tech)	年終配股 (year-end stock bonus)
電子新貴 (new high-tech stock millionaire)	網際網路 (Internet)	建廠地點 (new factory location) 待遇 (remuneration)
電路板 (circuit board)		產品價格 (product price)

Top half shows matches between expert and system or subject. The equivalent categories are marked with *.

the reverse is true on the sub-level (accepting H2b). On the top-level, the system might not match many of the humans' common-sense top-level categories, so the system performed poorly in precision compared to human subjects. We believe the reason might be that the domain-specific terms extracted by the system help more in the more specific sub-levels than in the more general top level.

For example, Table 4 shows that for a sub-level category, half of the categories selected by the system matched the experts' selections, whereas only about one-third of the categories selected by the subject matched the experts' selections. Therefore, the system precision was higher than the subject precision.

5. Evaluating visualization

While the previous section evaluated the quality of the knowledge map generated, this section presents an experiment investigating the effectiveness of the map interface in conveying the knowledge map to its users. The 1D alphabetically ordered subject hierarchy has a long history of delivering information having hierarchical relationships, but an entire list can become unmanageable as the size of hierarchy increases. In addition, the 1D display does not display information about semantic relationships among siblings. On the other hand, the 2D display of SOM not only presents semantic proximity through spatial proximity, but also utilizes visual cues such as size and color to deliver rich information about each category. For instance, the size of each category on a map indicates the number of documents within that category, while the colors specify how deep a user can go along a sub-tree.

However, the effectiveness of a hierarchical SOM was not supported by a previous empirical study that found users getting lost during browsing of noisy Internet web pages [8]. NewsMap, which was generated by using a much more precise and coherent news article testbed, adds a path indicator on the top of the map to specify the position of the map under investigation in the hierarchy. We hope such improvements, along with other visual cues such as color and size, show the differences between a 1D and a 2D SOM display in facilitating different information acquisition and evaluation tasks.

5.1. Experiment design and procedure

Previous human–computer interaction (HCI) research has provided a low-level, domain-independent taxonomy of tasks that users can perform when confronted with an interface, such as associate, cluster, compare, identify, locate, rank, etc. [39,43]. The "defeaturing" approach maps the domain-independent taxonomy of tasks into a specific domain. Such mapping ensures that only the visualization component will be evaluated. Because this approach has been shown to be valuable in evaluating a graphical interface [28], we adopted it to evaluate the interface of NewsMap instead of conducting a conventional usability test. We believe understanding the differences between the two interfaces in supporting different types of tasks may provide argument for selecting the NewsMap as an alternative interface. Based on the information and the functionality provided by the graphical interface, we selected three types of task whose definition and example tasks are presented in Table 5.

The experiment involved 20 subjects who are students from Taiwan. A subject completed two sessions: Finance News SOM vs. 1D display and Health News SOM vs. 1D display. Two sets of task were designed for each session and each task set contained three tasks to cover the three task types. During each session, a subject used one type of interface to finish a task set and used another interface to finish the second task set. The order of using interface type and the session order were randomly assigned to each subject. In addition, the task set was also randomly assigned to an interface type. During the experiment, subjects could take as long as they wanted to accomplish a task, but had to finish tasks one by one. Looking ahead was not allowed.

Since a subject could find correct answers using either interface type, task completion time alone was used to measure efficiency of the two interface types. During and after each session, a subject was encouraged to think aloud and gave feedback on the interface used.

5.2. Results and discussion

A one-way ANOVA test was run to compare the difference between the 1D and 2D displays and results were shown in Table 6. The two interface types were considered to have significant difference when the P value was smaller than 0.05. The experiment results were analyzed based on task types.

- *Identify* tasks required subject to search the hierarchy and browse the sub-categories of a category. Either of the interface type required only one click to display sub-categories of a category of interest. A subject then browsed the sub-categories over a list or a SOM map to obtain the answer. It appeared that this task was easy for both interface types and no significant difference was found between the 1D and 2D interfaces in the **identify** task.
- *Compare* tasks required a subject to do a sibling comparison. He/she needed to traverse the hierarchy to the correct level of comparison. The visual cues provided by the SOM map made it

Table 5
Definitions and example tasks of the three types selected

Task Type	Definition	*Sample experimental task*
Identify	Find a visual object based on certain attribute value.	*Which category is a sub-category of "insurance companies" (保險公司)?*
Compare	Compare based on certain attribute.	*Which of the following categories has more sub-levels? "semiconductor" (半導體) or "electronic stocks" (電子股) or equal.*
Associate	Form relationships between objects.	*Please list categories that are ancestors of "ChungHwa Telecom" (中華電訊).*

Table 6
Task completion time (in seconds) as measure of efficiency

Level	1D (s)	2D (s)	*p*-value
Identify	20.32	18.23	0.38
Compare	**63.85**	**28.62**	**0.00***
Associate	53.12	55.30	0.43

* Statistically significant.

easier for a subject to accomplish this task. The color shade of a category indicated how deep the sub-tree was, whereas it took a subject several clicks to figure out how many levels of the 1D interface he/she could go down. As a result, using a 2D interface was significantly more efficient in accomplishing the **compare** task than its 1D counterpart.

- *Associate* tasks asked a subject to identify the ancestor-descendent relationships among different nodes. This is more complicated than the other two task types. Both the 1D and 2D interfaces took the same number of clicks for a subject to get correct answer if he/she was on the correct path. However, it usually took several clicks for a subject to get on the right path. A subject experienced overload when scrolling back and forth on a 1D interface. At the same time, he/she had difficulty in remembering which sub-trees had been traversed when the SOM map was used. Overall, no significant difference was found between the two interface types in associate task.

Based on subjects' feedback and our observation, both interface types had their own advantages and disadvantages. Subjects liked the 1D display because they were accustomed to the folders arrangement through familiarity with the Microsoft Windows environment. It was easier to find a correct path systematically when using a 1D display. On the other hand, the 2D SOM map provided more visual cues and delivered richer information about each node within a hierarchy. It used spatial location for the semantic relationship among categories, size for the number of documents within a category, and color shades for the number of levels beneath a category. Those features enable easy sibling comparison. Based on the experiment results, it appears that the best strategy for using the NewsMap interface is to use the 1D display for the path management when traversing the hierarchy and to utilize the 2D SOM map to compare categories on the same level. The combination of 1D and 2D displays appears to be a promising approach to displaying knowledge structures with hierarchical relationships.

6. Conclusions and future directions

This research attempted to create high-quality hierarchical knowledge maps and to suggest effective map-based visualizations. We employed an automatic approach to generating hierarchical knowledge maps by using a statistical Chinese Indexer to represent news articles as a vector of phrases and a neural-network SOM Categorizer to reduce high dimensional vector space onto two-dimensional hierarchical knowledge maps. The first experiment showed that the categories generated by the system performed significantly better than the categories generated by human subjects in terms of recall and precision on the sub-level categories, but not on the top-level categories. The system was partially successful in capturing key indexing terms and reducing the high dimensional vector space representation into a two-dimensional map, thus overcoming the limitations of a human working memory. The second visualization experiment showed that meaningful visual cues in the display could reduce the time needed for task completion. For other tasks, both displays performed similarly. However, some users were used to 1D, while others preferred 2D. Our future research will try to incorporate the best features of one display into the other, so that a combination of both displays can better serve the users' needs in the future.

This research also represents the first step toward successfully extending this approach beyond the English-language to Chinese-language information analysis and visualization. Our language-independent indexing method combined with the generic SOM categorization technique potentially can be extended to analysis of many other knowledge sources in different languages. By relying on Java's Unicode capability, a new Java implementation of the PAT-Tree Phrase Extraction has been completed, and new research is under way to apply this approach to Spanish, Tamil and English to determine its external validity. Furthermore, a multilingual entity extraction system could be developed to support "business intelligence" by answering more targeted questions, such as "What are the major acquisitions of company X?" and "Who are the major competitors of company X?".

Acknowledgements

We wish to thank Yang Xiang, Yintak Lam, and Xue Wei for helping in the pilot study. Last but not least, special thanks go to Chienting Lin for being one of the domain experts and many helpful suggestions to the experiment as well as to the University of Arizona's Taiwanese Student Association for recruiting the experiment subjects.

This research was partly supported by NSF Digital Library Initiative-2, "High-performance Digital Library Systems: From Information Retrieval to Knowledge Management", IIS-9817473, April 1999–March 2002.

References

[1] P.G. Anick, S. Vaithyanathan, Exploiting clustering and phrases for context-based information retrieval, Proceedings of the ACM SIGIR Conference, Philadelphia, PA, (1997) 314–323.

[2] R.A. Baeza-Yates, G.H. Gonnet, Fast text searching for regular expressions or automaton searching on tries, Journal of the ACM 43 (6) (1996) 915–936.

[3] D.C. Blair, M.E. Maron, An evaluation of retrieval effectiveness for a full-text document-retrieval system, Communications of the ACM 28 (3) (1985) 289–299.

[4] T. Buzan, B. Buzan, The Mind Map Book, Penguin Group/ Dutton, New York, 1993.

[5] E. Carmel, S. Crawford, H. Chen, Browsing in hypertext: a cognitive study, IEEE Transactions on Systems, Man and Cybernetics 22 (5) (1992) 865–884.

[6] H. Chen, C. Schuffels, R. Owig, Internet categorization and search: a machine learning approach, Journal of Visual Communications and Image Representation Science 7 (1) (1996) 88–102.

[7] H. Chen, J. Martinez, D.T. Ng, B.R. Schatz, A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the worm community system, Journal of the American Society for Information Science 48 (1) (1997) 17–31.

[8] H. Chen, A.L. Houston, R.R. Sewell, B.R. Schatz, Internet browsing and searching: user evaluation of category map and concept space techniques, Journal of the American Society for Information Science 49 (7) (1998) 582–603.

[9] L.-F. Chien, PAT-Tree-based keyword extraction for Chinese information retrieval, Proceedings of the ACM SIGIR Conference, Philadelphia, PA, (1997) 50–58.

[10] Y. Choueka, T. Klein, E. Neuwitz, Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus, Journal for Literary and Linguistic Computing 4 (1983) 34–38.

[11] K. Church, P. Hanks, Word association norms, mutual information, and lexicography, Computational Linguistics 16 (1) (1990) 22–29.

[12] M. Dodge, NewsMaps: topographic mapping of information, Available at: http://mappa.mundi.net/maps/maps_015/ 2000.

[13] M. Dodge and R. Kitchin, Atlas of Cyberspace, Addison-Wesley (2001). (Also http://www.cybergeography.org/).

[14] J. Donath, K. Karahalios, F. Viegas, Visualizing conversation, Journal of Computer-Mediated Communication 4 (4) 1999.Available at: http://www.ascusc.org/jcmc.

[15] K.M. Drabenstott, M.S. Weller, The exact-display approach for online catalog subject searching, Information Processing & Management 32 (6) (1996) 719–745.

[16] Gartner Group, Summer Knowledge Management Workshop Report, 1998 (Summer).

[17] R.P. Holley, R.E. Killheffer, Is there an answer to the subject access crisis? Cataloging and Classification Quarterly 1 (2) (1982) 125–133.

[18] S. Kaski, T. Honkela, K. Lagus, T. Kohonen, WEBSOM-self-organizing maps of document collections, Neurocomputing 21 (1998) 101–117.

[19] J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon, J. Riedl, GroupLens: applying collaborative filtering to usenet news, Communications of the ACM 40 (3) (1997) 77–87.

[20] H. Lai, T.-C. Yang, A system architecture for intelligent browsing on the web, Decision Support Systems 28 (3) (2000) 219–239.

[21] X. Lin, D. Soergel, G. Marchionini, A self-organizing semantic map for information retrieval, Proceedings of the ACM SIGIR Conference, Chicago, IL, (1991) 262–269.

[22] R.P. Lippmann, An introduction to computing with neural networks, IEEE Acoustics, Speech, and Signal Processing Magazine 4 (2) (1987) 4–22.

[23] P. Maglio, R. Barrett, Intermediaries personalize information streams, Communications of the ACM 43 (8) (2000) 96–101.

[24] U. Manber, G. Myers, Suffix arrays: a new method for on-line string searches, SIAM Journal on Computing 22 (5) (1993) 935–948.

[25] G. Marchionini, An invitation to browse: designing full text systems for novice users, Canadian Journal of Information Science 12 (3) (1987) 69–79.

[26] G. Marchionini, B. Shneiderman, Finding facts vs. browsing knowledge in hypertext systems, IEEE Computer 21 (1) (1988) 70–79.

[27] M. Massicotte, Improved browsable displays for online subject access, Information Technology and Libraries 7 (4) (1988) 373–380.

[28] E. Morse, M. Lewis, Evaluating visualizations: using a taxonomic guide, International Journal of Human–Computer Studies 53 (5) (2000) 637–662.

[29] J.D. Novak, D.B. Gowin, Learning How to Learn, Cambridge Univ. Press, New York, 1984.

[30] Nua Internet Surveys, http://www.nua.ie (article ID 905356475) (2001).

[31] Nua Internet Surveys, http://www.nua.ie (article ID 905357873) (2002).

[32] T. Ong, H. Chen, Updateable PAT-Tree approach to Chinese key phrase extraction using mutual information: a linguistic foundation for knowledge management, Proceedings of the Second Asian Digital Library Conference, November 8–9, (1999) 63–84.

[33] R. Orwig, H. Chen, J.F. Nunamaker, A graphical, self-organizing approach to classifying electronic meeting output, Journal of the American Society for Information Science 48 (2) (1997) 157–170.

[34] D.G. Roussinov, H. Chen, Document clustering for electronic meetings: an experimental comparison of two techniques, Decision Support Systems 27 (1) (1999) 67–81.

[35] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

[36] B. Shneiderman, Creating creativity: user interfaces for supporting innovation, ACM Transactions on Computer–Human Interaction 7 (1) (2000) 114–138.

[37] B. Shneiderman, M. Wattenberg, Ordered TreeMap layouts, IEEE Symposium on Information Visualization, (2001) 73–78.

[38] F. Smadja, Retrieving collocations from text: Xtract, Computational Linguistics 19 (1) (1993) 143–178.

[39] S. Wehrend, C. Lewis, A problem-oriented classification of visualization techniques, Proceedings - IEEE Visualization, (1990) 139–143.

[40] R. Xiong, J. Donath, Creating data portraits for users, Proceedings of the 12th Annual ACM Symposium on User Interface Software and Technology, (1999) 37–44.

[41] C.C. Yang, J. Luk, S. Yung, J. Yen, Combination and boundary detection approaches on Chinese indexing, Journal of the American Society for Information Science, Special Issue on Digital Libraries 51 (4) (2000) 340–351.

[42] C.C. Yang, H. Chen, and K. Hong, Visualization of Large Category Map for Internet Browsing, Decision Support Systems, Special Issue on Web Retrieval and Mining, in press..

[43] M.X. Zhou, S.K. Feiner, Visual task characterization for automated visual discourse synthesis, Proceedings of ACM SIGCHI, (1998) 392–399.

Thian-Huat Ong is an assistant professor in the Management Information Science Department at the California State University at Sacramento. He was a member of the Artificial Intelligence Lab in the Management Information Systems Department at the University of Arizona. His research interests include multilingual information retrieval, data mining, knowledge management, and visualization. His research has appeared in *International Conference of Asian Digital Library*, *ACM/IEEE-CS Joint Conference on Digital Libraries*, and *Decision Support Systems*.

Dr. Hsinchun Chen is McClelland Professor of Management Information Systems at the University of Arizona and Andersen Consulting Professor of the Year (1999). He received the BS degree from the National Chiao-Tung University in Taiwan in 1981 and PhD degree in Information Systems from the New York University in USA in 1989. He is author of four books and more than 150 articles covering intelligence analysis, data/text/web mining, digital library, knowledge management, medical informatics, and Web computing. He serves on the editorial board of *Journal of the American Society for Information Science and Technology*, *ACM Transactions on Information Systems*, and *Decision Support Systems*.

Wai-Ki Sung is currently employed with Global Freight Exchange (GF-X). She received her master's degree in Management Information Systems from the University of Arizona and her bachelor's degree in Industrial and Manufacturing Systems Engineering from the University of Hong Kong. Prior to the current employment, she was a research assistant at the Artificial Intelligence Lab and a senior marketing specialist at IBM. Her research interests include e-commerce adoption and human computer interaction.

Bin Zhu received her PhD in Management Information Systems from University of Arizona. She is an assistant professor in the Information Systems Department at Boston University. Her current research interests include human–computer interaction, information visualization, computer-mediated communication, and knowledge management systems. She has been a lead author for papers that have appeared on the *Journal of the American Society for Information Science*, *IEEE Transaction on Image Processing*, and *D-Lib Magazine*. Her research also received an IBM faculty award in 2003.