Lin, Y., Brusilovsky, P., and He, D. (2011) Improving Self-Organizing Information Maps as Navigational Tools: A Semantic Approach. Online Information Review 35 (3), 401 - 424.

http://www.emeraldinsight.com/journals.htm?articleid=1937295&show=pdf

Abstract

- Purpose

  The goal of the research is to explore whether the use of higher-level semantic features can help us build better SOM representation as measured from a human-centered perspective. We also explore an automatic evaluation method that utilizes human expert knowledge encapsulated in the structure of traditional textbooks to determine map representation quality.

- Design/methodology/approach

  Two types of document representations involving semantic features have been explored: 1) using only one individual semantic feature, and 2) combining a semantic feature with keywords. A set of experiments were conducted to investigate the impact of semantic representation quality on the map. The experiments were performed on data collections that included single book corpus and multiple book corpus.

- Findings

  Combining keywords with certain semantic features achieves significant improvement of representation quality over the keywords-only approach in a relatively homogeneous single book corpus. Changing the ratios of the combined different features also affects the performance.
  While semantic mixtures can work well in single book corpus, they lose their increased effectiveness over keywords in the multiple-book corpus. This raises a concern about whether the semantic representations in the multiple book corpus are homogeneous and coherent enough to apply semantic features. The terminology issue among textbooks negatively impacts the ability of the SOM to generate a high quality map for heterogeneous collections.

- Originality/value

  We explored the use of higher-level document representation features for the development of better-quality SOM. In addition, we piloted a specific method for evaluating the SOM quality based on the organization of information content in the map.

1. Introduction

Information maps (Kohonen, 1982) are becoming popular as interfaces to view and access large data collections such as digital libraries (DL). Unlike traditional search-based access, which provides selective and fragmented access to information, information maps allow users to comprehend large collections, to focus on the most interesting parts, and to explore specific resources in the context of their relationships to other resources and the library. Properties of information maps make them an excellent complement to search and browsing interfaces for DL. A recent study comparing student use of search, browsing and information map interfaces in an educational DL (Brusilovsky et al., 2005) found that information maps were the method most preferred by students for accessing information; they were four times more popular than traditional search-based access methods. Several kinds of maps have been explored as interfaces to access large collections of resources (Börner and Chen, 2002, Yang et al., 2003, Dang et al., 2009, Perugini et al., 2004). Among these approaches, A self-organizing Map (SOM) (Kohonen, 1982) is frequently considered to be the most promising mapping approach for large document collections. While being most popular as a tool for two-dimensional clustering in engineering science, medicine, biology, and economics (Kohonen, 1998, Oja et al., 2003), SOM is becoming increasingly popular in producing information maps which support user navigation (Brusilovsky and Rizzo, 2002, Chen et al., 1998, Dang et al., 2009, Lin et al., 1991, Rauber and Merkl, 1999, Roussinov and Chen, 1998, Yang et al., 2003). SOM clusters similar resources into the same cell or nearby cells on the map, so that users will be able to easily identify the relatedness of the categories created based on spatial proximity. In comparison with other mapping techniques, the SOM technique is a simple, straightforward, and

highly-scalable random projection method suitable for any size collection of items. It does not require explicit connections between documents or the presence of any kind of metadata.

On the other hand, previous researches (Brusilovsky and Rizzo, 2002 , Chen et al., 1998) have indicated that the artificial organization produced by SOM may not be easily understood by all users. Users are frequently unsure about the reason why a specific combination of resources was placed in the same cell, resulting in a negative experience when navigating through SOM. The main challenge of our research was to produce a SOM which would provide a closer match to a person's conceptualization in a specific domain. We hypothesize that a potential reason for the lack of human-friendly organization of SOM is the keyword-level document representation, which is currently used to construct the maps and to represent the contents of the cells in the maps. Simple keyword representations are known to have several shortcomings at the semantic level. Several studies in the area of information retrieval have indicated that replacing or augmenting simple keywords with semantically-richer features such as noun phrases or concepts could lead to significant performance improvement in certain domains (Gonzalo et al., 1998, Stokoe et al., 2003). Semantic representations have been used to solve the challenge of traditional keyword-level representation in diverse applications such as information retrieval (Basile et al., 2008), the heterogeneous Web (Tang, 2002), and question-answering systems (Vicedo and Ferrández, 2000).

Expecting that a similar approach can help us produce better quality information maps, we explore the integration of several semantically-rich features alone and in combination with keywords for map construction. Therefore, the first research topic examines how to enhance SOM quality using semantic features in the SOM construction.

Any research focused on producing better information maps for end users should start by defining a meaningful approach to measure this quality. However, only a few studies have focused on SOM quality issues, and these studies were concerned mainly with the quality of the clustering algorithm and techniques for its application (Lo and Bavarian, 1991, Kiang et al., 2006, Su et al., 2002). While a number of studies have focused on the navigational use of SOM by human users (Brusilovsky and Rizzo, 2002, Lin et al., 1991, Rauber and Merkl, 1999, Roussinov and Chen, 1998), there were no approaches suggested to evaluate map representation from a human-centered point of view. This caused us to pay special attention to the evaluation of the map representation quality from a human perspective.

The remainder of the paper is organized as follows: in Section 2, a literature review surveys research relevant to Self-Organizing Maps. Section 3 discusses the goal and research questions of this study. Section 4 introduces our main innovation, the semantic approach to SOM construction. The context of our research and the Knowledge Sea information mapping system utilized for the study are introduced in this section, as well. The other innovation, the "textbook" method of SOM evaluation, is proposed in Section 5. Section 6 presents the results of our studies, which compare the quality of SOM produced with the use of different features and their combinations. Finally, our conclusions are discussed in Section 7.

2. Self-organizing map (SOM)

The self-organizing map (SOM) tool is a type of an unsupervised neural network model developed by Teuvo Kohonen (1982). SOM has the ability to reduce the dimensions of data by applying self-organizing neural networks (Kohonen, 1998). Each neuron, a processing unit in SOM, is associated with a weight vector and is positioned on a map. During the learning

stage, as the weights of each unit change, their corresponding positions on the map would change and consequently move the input points to a different location. After the iterative learning stage, the movement caused by weight change becomes slower and the units become more stable in the input space.

The most attractive characteristic of SOM is the ability to transform a high-dimensional input space into a two-dimensional output space which faithfully preserves the structure of the input data. SOM has spread into numerous fields as a research methodology, particularly in analyzing large volumes of high-dimensional data. The SOM literature can be organized into two branches. One focuses on the study of the relationships between the topical categories. Schatz‘s study (1996) showed that SOM has been adopted by many academic projects for textual document classification. Oja categorized human endogenous retrovirus (HERVs) into meaningful groups using SOM (Oja et al., 2004). Dina (2005) explored automatic document categorization methods by comparing SOM and Learning Vector Quantization (LVQ). The other branch focuses on an interface for browsing and searching diverse collections. Lin (1991) was the pioneer of using SOM as a tool for information access. Roussinov (1998) proposed a multi-level SOM, extending a group of cells into a second layer to assist users with navigating through a large corpus. Rauber and Merkl (1999) showed that the LabelSOM method of automatically labeling the various topical clusters found in the map offered an instant overview for users. Brusilovsky and Rizzo (2002) applied SOM to develop a landmark-based navigation system, Knowledge Sea, to provide access to a large collection of educational resources. Chen and colleagues have explored the use of multi-level SOM for information access in several practical domains (Dang et al., 2009, Yang et al., 2003).

Meanwhile, several different approaches have been proposed to improve the SOM algorithm in order to create better maps. Lo (1991) focused on the selection of neighborhood function, and Kiang (2006) proposed a circular training algorithm to overcome the "boundary" effect on topological representations. In another study, an incremental learning algorithm was applied (Jun et al., 1993). Su (2002) launched an efficient initialization scheme to construct an initial map and eventually generate a better-performing map.

Works focusing on SOM representation quality measures are rare in the literature. Most publications have been mainly concerned with exploring energy functions to improve the quality of map topology (Erwin et al., 1992, Heskes, 1999). Kaski (1996) and Pölzlbauer (2004) compared existing methods for quantifying the quality of SOM. However, most publications were concerned with topological improvement, not with the quality of SOM as a tool for information access. In light of the increase in SOM applications designed for navigational use, such as multiple-layered SOM (Roussinov and Chen, 1998) or an incremental SOM (Benabdeslem and Bennani, 2004), it is critical to develop an evaluation method which is centered on the quality of maps from a human-centered point of view. For most users, it is preferable that related contents are grouped together and relationships are easily identifiable, so that searching and browsing can be supported. Therefore, this paper primarily investigates the ability of SOM to organize the content so that it is arranged similarly to the human approach to content organization.

3. Research Questions

The goal of this research is to determine whether or not the use of higher-level semantic features can help to build better SOM representation as measured from a human-centered perspective. Since the higher-level features can be used to produce SOM in two ways (in place of - or in addition to - traditional keywords) the following research questions are addressed.

Q1: Can we produce better SOM by replacing keyword-level document representation with semantic-level representation?

Q2: Can we improve the quality of SOM by enhancing keyword-level document representation with semantic features; if so, which combination of features – in which ratios – would produce the best map?

4. The Semantic Approach for SOM Construction

The problem of building SOM using semantic features can be explored in several different contexts. Each context defines a specific combination of a domain and a version of the SOM approach used to organize documents within this domain. Chen et al. (2003) used a clustered hierarchical SOM to provide access to a large volume of medical information. Brusilovsky and Rizzo (2002) applied a traditional one-level SOM to provide access to multiple electronic textbooks. Defining a context for this kind of research is very important: the domain delineates the choice of specific semantic features while the applied map construction approach defines how these features can be used to build the SOM (instead of or in parallel with traditional keywords). In addition, a clear understanding of the research context helps us comprehend the problem and evaluate possible solutions. As an example of this importance, this section starts with a description of the context that instigated our research.

After that, we discuss several semantic features available in the selected domain; we also outline our approach to use these features for SOM construction.

4.1 The Context

Our research is directly motivated by our experience with **Knowledge Sea** (Brusilovsky and Rizzo, 2002), an integrated system for accessing educational resources. In the context of Knowledge Sea, a SOM-based information map was used as one of three key approaches (in addition to browsing and search) to access a collection of educational resources (tutorials, books, and handouts). Since 2002, several versions of Knowledge Sea were tested in many undergraduate and graduate classes in three domains: C-programming, information retrieval, and human-computer interaction (Farzan and Brusilovsky, 2005, Brusilovsky et al., 2004, Brusilovsky and Rizzo, 2002). The SOM-based information map in Knowledge Sea is a two-dimensional array of cells arranged 8 by 8 (Figure 1). Each cell displays a set of keywords and *landmarks* using features, different icons and background colors. The landmarks provide additional navigation support to help users locate the cells which contain the most relevant documents. The icons and background provides additional navigation cues. By clicking on a cell, users can access documents belonging to the cell along with a list of the most relevant keywords associated with the cell's content and a navigation map indicating the position of the cell within the whole map. The properties of SOM ensure that the degree of similarity between the documents is reflected in the proximity in locations on the map. The most similar documents are located in the same cell, the slightly less similar in the adjacent cells, and so forth. On the map level, the distance between cells reflects the similarity between documents grouped within these cells. Therefore, we could utilize

Knowledge Sea as our platform to test SOM quality by evaluating whether relevant textbook documents are assigned to cells adjacent to each other on the map.

Knowledge Sea proved to be a useful information access tool in an educational context. The log analysis demonstrated that the map emerged as the most popular tool to access electronic textbooks, outperforming search and browsing (Brusilovsky and Rizzo, 2002). Students also ranked the map highly in several rounds of classroom studies. Yet, our interviews with students and some unsolicited comments indicated that students are sometimes puzzled by the placement of specific documents in the map. More specifically, it was confusing that conceptually-similar documents such as subsequent sections of the same book were located far away from each other on the map. It was exactly this experience that motivated the work presented below.

4.2 The Domain and the Semantic Features

The choice of the domain is critical because it defines the kind of semantic features available for using in SOM construction. For a SOM in a medical domain such as (Chen et al., 2003), noun phrases could be the appropriate semantic feature; however, for a SOM which provides access to news magazines (Rauber and Merkl, 1999), it would be more appropriate to select named entities (i.e. names of people, places, or things). In our context, the domain is a set of electronic textbooks and similar sources focusing on teaching a specific subject. Here, the most natural semantic features are domain concepts, which are presented and explained to the students in these textbooks. These concepts have to be either extracted from the text or provided by experts. For the extraction option, we tried two state-of-the-art approaches: noun phrases and

Yahoo! concepts. For expert-provided concepts, we explored glossary terms containing concepts that are specific to our textbook context. Below, we provide a more detailed discussion of the semantic features used in our study as compared to traditional keywords.

A **keyword** is defined as a single term with special significance in the textbook corpus. Standard parsing and tokenization methods were used to retrieve keywords from the corpus. The Porter Stemming Algorithm was performed and we also created approximately 150 stop words to filter out non-related keywords.

A **noun phrase** in our study refers to a chunk of text, which is identified using a particular language processing tool. The phrase structure is assumed to consist of its root (which is a noun or a pronoun) and, possibly, modifiers. We used the Arizona Noun Phraser (Leroy and Chen, 2005, Tolle and Chen, 2000) to extract all noun phrases from the corpus. Then, stop words were removed to generate a meaningful phrase list. The Noun Phraser is based on a part-of-speech tagger (Brill, 1993) and noun phrase identification rules from NPtool (Voutilainen, 1993), a commercial noun phrase extractor. The purpose of using these noun phrases is to verify that multiple words often offer a more precise meaning than single words; therefore, they can help to reduce ambiguities in text (Harper, 1992). In our studies, noun phrases were considered as the lowest level semantic features (after keywords). In fact, a significant fraction of extracted noun phrases was as effective as keywords extracted by a regular keyword extraction process.

A **concept** in our study refers to a significant word or phrase in the corpus identified by the Yahoo Term Extraction Web Service[1]. The service has been used for a variety of different purposes. For instance, Y!Q[2] uses it to determine key

1. http://developer.yahoo.com/search/content/V1/termExtraction.html

concepts within the search context and to apply those concepts to augment a user's search query. Comparing with the noun phrases mentioned above, concepts can be considered as higher-level semantic features. Their extraction is based on more sophisticated approaches to text analysis than the historically older and simpler noun phrasing techniques.

A **glossary term** is an important domain concept, which was selected for inclusion in the textbook glossary and extended with a clear definition. By their nature, glossary terms are the highest-level semantic features. Glossary terms are manually designated by the authors who are domain experts as key concepts of the domain, as opposed to other automatically extracted features (keywords, noun phrases, and concepts). At the same time, a set of glossary terms is not as comprehensive as automatically-extracted semantic features since humans are typically selective in picking a set of terms for the glossary

Our specific interest in exploring glossary terms causes us to select a digital library of textbooks on information retrieval (IR) for our study. From among several domains, which were prepared for Knowledge Sea mapping, this collection has the largest number of glossary terms. This library contains the full content of four classic textbooks in the IR field: “Finding Out About” (Belew, 2000), “Modern Information Retrieval” (Baeza-Yates and Ribeiro-Neto, 1999), “Information Retrieval” (Van Rijsbergen, 1979), and “Information Storage and Retrieval” (Korfhage, 1997). From a SOM point of view, each subsection of each textbook is considered to be a separate document. In total, there are 714 documents in the library. The glossary sections of these textbooks contain 402 unique glossary terms.

---

2. http://yq.search.yahoo.com/

4.3 Semantic map generation

4.3.1 The Traditional Approach: generating SOM using keyword-level document representation

In general, using SOM to generate an information map involves two steps. The first step is feature extraction. In the case of using keywords to represent documents, keywords from the corpus are extracted and selected using standard IR keyword identification and weighting techniques. Once the selected keywords are defined, each document in the corpus has its corresponding vector representation. The second step is map generation and document assignment. The map size is often predefined as an m-by-n matrix that contains m*n cells (m, n: the number of cells). Each cell is represented by a vector in the same space as the document vectors. Therefore, with a preselected similarity measure such as the cosine similarity or neural network techniques, documents can (one-by-one) be inserted into the map nearest to the most similar cell. The distance between the cells represents the level of relatedness among the vectors in these cells. The closer the relationship between vectors of features, the closer the geographic positions will be.

4.3.2 The Semantic Approach: Generating SOM with semantic features

When the documents are represented by semantic features rather than keywords, both the feature selection and the map generation steps are essentially identical to that of using keywords. Technically, the only difference was the feature extraction process. Semantic features were extracted from the corpus with special tools. Noun phrases were identified and extracted using the Arizona Noun Phraser; concepts were identified and extracted using Yahoo concept extractor; glossary terms were identified by the book authors (who created corresponding glossaries) and were extracted by a simple script. After that, the extracted semantic features were processed in a standard way to produce a representation of every document as a weighted

vector of semantic features. In total, for each kind of features, we obtained an independent set of vectors representing the original documents. To produce single-feature maps, we used the corresponding set of vectors in the same way as keyword vectors are used in map generation and document assignment processes (Figure 3). Generating document representation and SOMs using mixtures of keywords and various semantic features was done in a slightly more sophisticated way, which is presented in detail in Section 6.3.

4.3.3 Control parameter settings

According to previous discussions on SOM (Su et al., 2002, Kohonen, 1990), map generation can be affected by different parameters such as learning iteration, learning rate, and neighborhood size. In order to achieve comparable results in our map evaluation, we relied on heuristic rules to determine these parameters prior to beginning the experiment. First, according to the literature, the number of iterations should be at least 500 times larger than the number of neurons (Kohonen, 1990). However, too many iterations may cause the problem of overfitting while generating SOM. Therefore, based on a pilot experiment we conducted on the document collections, we set the number of iterations at 2000, since any value larger than 2000 produced almost identical maps and over-representation and under-representation were not issues at 2000 times.

The map size is defined as 8-by-8 to be consistent with Knowledge Sea. We experimented with three neighborhood sizes (2, 3, and 4) with ten different learning rates ranging from 0.1 to 0.01. Eventually, through a manual checking of generated maps, we defined the neighborhood value as 2 and the learning rate as 0.1.

Even with all of the aforementioned parameters being pre-selected, the result of the map generation was still not determined because SOM is random by the nature. To avoid any side-effects caused by semi-sorted inputs, SOM selects

random seeds in the initialization of the algorithm inputs (Amarasiri et al., 2006). Differences in random seeds could cause the generated map to have different topological orders (Kohonen, 1998), which results in the overall performance of a particular map being strongly related to random components. To compensate for the uncertainty in a single random seed, we generated ten maps for each domain representation using ten random seeds, and averaged the results.

5. The Evaluation Approach

Evaluating the quality of SOM from a human-centered navigational point of view is a challenging issue that has not been studied thoroughly. As mentioned earlier, existing approaches to SOM evaluation do not take into account human perspective. A commonly-used methodology for user-centered quality evaluation is to apply expert judgments or to conduct user studies. Although they are potentially useful to identify the quality of SOM, these approaches are limited in some respects, such as budget, domain knowledge, subjective bias, and unrepeatable results. The main problem here is the nature of the SOM approach, which is determined not only by the original vectors and features, but also by several generation parameters such as random seed or learning rate. Even with key parameters fixed, we had to generate ten maps for each approach using different random seeds and had to compare two sets of maps, rather than simply comparing two individual maps. Comparing such a large number of maps in a user study is not feasible particularly as map quality is difficult for users to judge. In fact, it is not easy to evaluate the quality of even a single map, as a subject would need to examine every cell in an attempt to rate the similarity of the resources in the cell from a human "conceptual" point of view. Thus, we cannot rely on

traditional user studies, but have to rely on some form of "encapsulated" human judgment to evaluate a large number of maps.

In searching for this encapsulated human judgment, we turned to human expert knowledge encapsulated in the structure of traditional textbooks. We believe that similarities between concepts are encapsulated in a textbook's typical structure. Moreover, it is not simply a random user judgment (as we can get from a user study); it is a judgment from experts in the field. These considerations defined our evaluation approach. To explore whether higher-level semantic features can produce more "human" SOMs, we used a collection of well-structured textbooks as the corpus for the study and used the structure of these textbooks as an alternative gold standard to evaluate the quality of SOM. This approach is explained in the next section.

5.1 A Textbook-Centered Evaluation Approach

The textbook-centered evaluation approach, which we propose, is based on properties of academic textbooks. By their design, textbooks focus on a specific issue (a topic) about the domain in each chapter. Within a chapter (first level), more specific concepts related to the chapter's key issue are systematically examined section-by-section (second level), with each section devoted to a specific set of concepts. Each third level subsection (if a specific book goes down to the third level) typically examines an even smaller, yet, consistent set of concepts. However, by the nature of their being grouped in the same section, we expect some reasonable conceptual overlap between subsections of the same section and still some better-than-average overlap between sections of the same chapter. As the association of concepts is understood, it will be easy to identify whether the deployment of concepts in the knowledge map is consistent with the organization of the domain.

In this study, we defined a cluster as a section of a chapter in a textbook. The assumption is that a more human-centered SOM construction approach, the one which better preserves the conceptual structure of the domain as identified by the human expert, should place documents belonging to the same conceptual cluster closer to each other on the map. In order to avoid some outlying topics and sections in a chapter introduction that might not exactly represent the concepts in the document, our study only considered the third-level sections as documents.

For instance, in Figure 4 Section1-1-1 is conceptually close to Section1-1-2, Section1-1-3, and Section1-1-4 but quite distant from Section4-3-1 (section1-2-3 means chapter 1, section 2, subsection 3). Thus, a good map should display Section1-1-1 and Section1-1-2 closer together than Section1-1-1 and Section4-3-1. Figure 5 places Section1-1-1 closer to Section4-3-1 than to Section1-1-2 or Section1-1-3, which may indicate a conceptual problem with this map.

In this study, we assessed the map quality by calculating the average corpus spread in 4 steps:

1. The spread of two documents is defined by the Euclidean distance (Teknomo) between the cells that documents D1 and D2 are in. (X and Y)

   $$Lp(D1,D2) = \sqrt{\sum_i (d1i - d2i)^2}$$ i = X dimension, Y dimension

   (1)

   For example, if Section1-1-1 is located in cell (0,0) and Section1-1-2 is in cell (0,2), their spread is 2.

2. The spread of one *cluster* (Sc, a set of third-level subsections belonging to the same second-level section such as 1-1-1, 1-1-2,and 1-1-3) is defined as the average spread of all document pairs in the cluster

$$Sc = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}[\frac{Lp(Di,Dj)}{2}]}{n!/(n-2)!2!}$$ n = # of documents in the cluster (2)

3. The spread of a whole corpus (*Sb*) is the average of its clusters

$$Sb = \frac{\sum_{i=1}^{n} Sci}{n}$$ n= # of clusters (3)

5.2 Feasibility Examination

Our evaluation approach is based on the assumption that documents within a textbook cluster (i.e., subsections in the same section) are more similar to each other than to documents outside of that cluster. To check whether this assumption is defendable, we separately calculated average keyword-based cosine similarity between documents within each cluster and across different clusters. Table 1 shows that for each of the four books used in our study, subsections belonging to the same cluster are much more similar to each other than to subsections from different clusters. The Wilcoxon Signed Ranks shows that this difference is significant ($p < .001$) for each of the books. This provides some reasonable evidence that documents within a particular cluster are more similar to each other than to documents found outside of that cluster.

Table 1: The comparison between the cross-cluster and the within-cluster

Book	Cross-clusters Average Similarity	Within-clusters Average Similarity
(Belew, 2000)	.1	.28
(Baeza-Yates and Ribeiro-Neto, 1999)	.08	.35
(Van Rijsbergen, 1979)	.07	.38
(Korfhage, 1997)	.14	.40

6. The Study

Motivated by the research questions presented in Section 3, we conducted a set of experiments to investigate the impact of semantic representations on the map quality. Two hypotheses were examined.

H1: The semantic representations would provide a higher-quality map than keywords.

H2: Combining keywords with certain semantic features would achieve significant improvement in map quality over the keywords-only approach.

The experiments were performed on data collections that included only one individual book (single book corpus) or all four books (multiple book corpus). To find answers to both research questions, the experiments also examined two types of document representations involving semantic features: 1) using only one individual semantic feature, and 2) mixing a semantic feature with keywords.

Therefore, the experiments were modeled as several ANOVA experiments. The dependent variable is the spread of the corpus (*Sb*), which indicates the map quality. The independent variables include corpora, features (keyword, noun phrase, concept, and glossary), and feature mixtures (the combination ratio of features and the combination weights of features).

In the experiments, the four types of features were extracted from each of the four books individually. For each type of feature, we generated 10 SOMs based on a constant set of random seeds. Documents were then assigned to each map, and the final map was then evaluated based on its spread of a cluster (Sc). To assess the performance of each type of feature, we considered mean, median, and minimum *Sb* calculated for each of the 10 maps generated using the feature.

6.1 Individual Feature Analysis in a Four-Book Corpus

In this part of the experiment, we used all four books to generate maps. We were interested in comparing the spread (*Sb*) for the four individual representations (i.e., how far apart a map based on each kind of feature spreads textbook sections from the same cluster). When the representations were keywords, noun phrases, or concepts, the top 600 features selected based on their weights were extracted from the corpus individually. As for glossary terms, only 402 terms were extracted which represented the total volume of the glossary collection.

As shown in Table 2, contrary to the expectation that those semantic features (noun phrases, concepts, or glossary terms) would generate higher quality maps than maps based on keywords, the mean of the spread (*Sb*) for keywords with ten random seeds has the lowest value (1.79) and also produces the minimum value (1.55) among all of the results. This pattern is also found when looking at the lowest mean of the spread, the lowest median, and the minimum value among all results produced by keyword (Table 2).

Table 2: The spread (*Sb*) of the multiple book corpus with a constant set of ten random seeds

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	Min	Mean	SD	Median
Keywords	2.07	1.94	1.60	1.85	2.01	1.78	1.86	1.71	1.55	1.55	1.55	1.79	0.189	1.81
Phrases	2.15	1.88	1.93	1.83	1.64	1.76	1.73	2.23	1.74	2.03	1.64	1.89	0.193	1.86
Concepts	2.23	1.95	1.82	2.30	2.03	2.01	2.13	1.80	2.30	1.83	1.80	2.04	0.194	2.02
Glossary	2.56	2.65	2.57	2.54	2.79	2.89	2.85	2.75	3.05	2.78	2.54	2.74	0.165	2.76

*R1, R2, R3~R10 indicate 10 different random seeds

The ANOVA results show that there is a significant difference among the features, $p<.001$. The mean of the spread (*Sb*) for keywords with ten random seeds is significantly lower than that for concepts, $p=.002$, and glossary terms, $p<.001$

(Figure 6). The results do not support our hypothesis that semantic representations would provide a higher-quality map than keywords. These results certainly demonstrate the need for further investigation of the initial premise.

One possible source of the negative results could be the fact that the books in the collection are still too heterogeneous. Although the four books we used are all textbooks on information retrieval, each book still has reasonably distinctive terms to express the concepts in this domain. We noticed this issue while analyzing and merging the glossaries from the books. These glossaries are substantially different from each other, with almost no overlap. Out of 402 glossary terms extracted from the four glossaries, only 9 terms appear in more than two books and none of the terms appear in three or more (Table 3).

Table 3: Indexing heterogeneity for different features

# of features shared by	Keywords	Noun Phrases	Concepts	Glossary terms
1book	0	47	307	393
2 books	9	100	117	9
3 books	88	154	114	0
All 4 books	503	299	62	0
Total	600	600	600	402

While this may look strange, this result stems from the nature of terms included in a glossary - highly specific and complex domain terms, which require explanation. With this level of complexity and specificity, even two books in the same domain frequently use slightly different representations for the same concept. For example, “Finding Out About” (Belew, 2000) uses “relevance” whereas “Modern Information Retrieval” (Baeza-Yates and Ribeiro-Neto, 1999) employs “user relevance” to represent the same concept in glossaries.

If the sets of features used to index the different books in the collection are essentially different, such indexing can be called heterogeneous. In contrast, if these sets are very similar, such indexing can be called homogeneous. Further analysis of Table 3 demonstrates that the quality of the information map produced with a specific kind of feature decreases with the

increase of heterogeneity of indexing using this kind of feature. As we can see, indexing with glossary terms is most heterogeneous: the set of glossary terms used to index different books has almost no overlap. Switching from highly specific manually-selected glossary terms to less specific automatically-extracted concepts decreases the heterogeneity of source representation (62 concepts were found in all four books!) and increases the quality of the information map. On the other end of the spectrum, simple keyword indexing provides the most homogeneous representation (503 keywords were found in all four books!) and the best map. Noun phrases fall between concepts and keywords, being apparently more specific than keywords, yet less specific than concepts. To investigate whether the heterogeneity was really the main source of the observed decline in quality, we decided to explore the performance of different kinds of features when building SOM for a single-book domain, which apparently offers higher homogeneity of representation.

6.2 Individual Feature Analysis in a Single Book Corpus

In view of the terminology issue, comparing the performance of different features in generating a map for a single book became a focal point in the study. "Modern Information Retrieval" (Baeza-Yates and Ribeiro-Neto, 1999) was the largest book in our corpus, containing 15 chapters, 308 sections, and 154 glossary terms (the largest glossary among the four books). Therefore, this book was selected to be the corpus in the single book study. The process of map generation, document assignment, and distance comparison was identical to the experiments using the four-book corpus.

Table 4: The spread (*Sb*) of the single book corpus with a constant set of ten random seeds

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	Min	Mean	SD	Median
Keywords	1.71	1.78	1.71	2.00	1.84	1.80	1.79	1.51	1.74	1.68	1.51	1.75	0.125	1.74
Phrases	1.90	1.84	1.65	1.64	1.74	1.50	1.78	1.63	1.82	1.87	1.50	1.74	0.129	1.76
Concepts	1.95	1.71	2.10	1.93	1.96	1.76	1.77	2.16	1.87	2.35	1.71	1.96	0.199	1.94

Glossary	2.49	2.63	2.31	2.32	2.12	2.31	2.50	2.45	2.35	2.30	2.12	2.38	0.142	2.34

*R1,R2,R3~R10 indicates 10 different random seeds

The results show that the mean of the spread (*Sb*) for phrases with ten random seeds has the lowest value and also produces the minimum value among all results (Table 4). However, according the ANOVA results, the mean of the spread (*Sb*) for phrase features is not significantly different from that for keyword features, $p = .997$. The analysis found that our hypothesis that higher-level features perform better than the classic keyword feature within a single book corpus is still not supported. Nothing outperforms keywords, although phrases performed equally well. The performances of concept, $p = .022$, and glossary, $p < .001$ features are still significantly worse than the performance of keywords (Figure 7).

Poor performance by concepts and glossary items in a single book corpus demonstrates that the heterogeneity may not be the most critical difference between indexing with higher-level features or with traditional keywords. To study the problem further, we compared low-level differences between several kinds of indexing. Most interesting is the issue of indexing density: how many features of different levels can be found on a single page and, vice versa, how many pages are indexed by the same feature? Our analysis revealed essential differences in indexing density between all four kinds of features: once we moved from very generic keywords to highly-specific glossary terms, the indexing density falls rapidly (Table 5). In keyword-level indexing, each book section is represented by 600 high frequency keywords, with 77.88 unique keywords per page and almost 200 recognized keywords overall. On the other end of the spectrum, each section is represented on average by only 6.13 unique glossary terms. Noun phrases are very close to keywords (most of them being, in fact, single nouns) while concepts stand somewhere between the two extremes. The low density of indexing clearly shows that both concepts and glossary items, when used alone, are not able to represent the content of the pages sufficiently well.

While each concept or glossary term can represent some aspect of page meaning on a deeper level, the low number of concepts or items per page means that this representation may be "patchy"; i.e., some aspects of the page content will not be represented at all. This fact is also confirmed by the significant increase in the number of pages, which have none of the features listed in the top 600 concepts or the top 402 glossary items (Table 5). It is interesting to observe that the performance of higher-level features (Table 4) does not degrade as rapidly as indexing density (Table 5). Thus, we can speculate that the increased "depth" of indexing with higher-level features could positively affect the quality of the maps, but it still can't compensate for the rapid fall of indexing density and the resulting "patchy" representation of units.

Table 5: Density of indexing with different kinds of features

	keyword	noun phrase	concept	glossary
Average term length (in words)	1.000	1.003	1.340	1.873
Average number of features per unit	191.15	142.34	55.68	15.50
Average number of unique features per unit	77.88	60.94	23.85	6.13
Average number of units per feature	92.69	72.52	28.39	10.89
Units with no features	0.000	1.000	5.000	53.000

One potential way to increase the density of indexing while maintaining the semantic depth of representation can be a radical increase in the number of features used for indexing (i.e., from 600 top features to 2000 or more). However, this approach will also decrease the speed of map construction and will not work with glossary items since there are only 402 such terms. Thus, in our study, we decided to explore an alternative approach: mixing keywords and higher-level features when indexing the documents; for example, using the top 300 keywords and the top 300 concepts, We expected that the presence of concepts in such a mixture would allow us to represent the most critical aspects of unit meaning at a deeper level, while the presence of keywords would allow a high-level of indexing density to be maintained and to avoid "patchy" representation of a unit's content. The research question then became whether higher-level semantic features could be merged

with the classic keywords to improve the quality of a map, and if so, which mixture of these features would provide the best results.

6.3 Feature Combination Analysis in a Single Book Corpus

Tomuro (2002) investigated whether or not semantic features could enhance classifying questions by comparing two feature sets: one with lexical features only, and the other with a mixture of lexical and semantic features. The study's purpose was quite similar to ours in this research. Therefore, after investigating the performance of individual features, this section explores combining keywords with other features to enhance performance. Two approaches are applied: one is a mixture based on different combination ratios of the features, and the other is focused on adjusting the weights of the features.

6.3.1 Adjusting feature ratio

In the individual feature analysis, keywords showed the greatest potential in both corpora. Therefore, in order to obtain comprehensible semantic representations, three higher-level features were paired with the keywords, producing three types of mixtures: 1) keyword and phrase, 2) keyword and concept, and 3) keyword and glossary term. The study assessed these mixtures individually and evaluated the patterns of the mixtures in single book corpus. Keeping the total number of features constant, we explored 5 different ratio combinations: keyword-only, 80% keyword and 20% target feature, 50% keyword and 50% target feature, 20% keyword and 80% target feature, and target feature-only. For example, the keyword-only combination had 600 keywords whereas the 80% keyword and 20% target feature combination had 480 keywords and 120 target feature terms (Table 6). The whole process of generating 10 maps, section assignment, and distance calculations was

performed for each of these combinations. The ANOVA results show that the combination of keyword and phrases is not able to outperform keywords significantly.

As Table 6 shows, the use of feature mixtures does affect the quality of the resulting SOM. For each of the three higher-level features, there is at least one combination which produces better results than single keywords. More importantly, we find a significant difference between keyword-only and any other keyword/concept combination in the single book corpus, *p=.005* (Figure 8). In fact, any keyword/concept combinations performed better than keywords alone. In addition, we observe that slightly better results are achieved when the keyword ratio is set as the higher of the two ratios in the combinations.

Table 6: Means and SDs of *Sb* by corpus*target feature*mixture type in single book corpus

Corpus	target feature	mixture type	Mean	Std. Deviation	N
Single	phrase	1 k	1.75	.1248	10
		0.8k+0.2p	1.67	.1686	10
		0.5k+0.5p	1.78	.2136	10
		0.2k+0.8p	1.76	.1562	10
		1p	1.74	.1283	10
	concept	1 k	1.75	.1248	10
		0.8k+0.2c	**1.53 ****	**.1915**	**10**
		0.5k+0.5c	**1.56 ****	**.1990**	**10**
		0.2k+0.8c	**1.65 ****	**.1413**	**10**
		1c	1.96	.1998	10
	glossary	1k	1.75	.1248	10
		0.8k+0.2g	1.70	.1559	10
		1g	2.38	.1423	10

** significant at $p<.01$
N = number of maps

Next, the study moved on to compare keyword-only with each combination, and to look for the best ratio of keyword and concept mixing in the single book corpus. The marginal comparisons reveal that the keyword-only approach has a significantly larger mean of the spread (*Sb*) than the combination of 80% keyword and 20% concept, $p = .004$, or the combination of 50% keyword and 50% concept, $p = .009$ (Table 6). Even though the combination of 80% keyword and 20%

concept has a lower mean of the spread (*Sb*), which means it performs better than the combination of 50% keyword and 50% concept, the combination with a higher percentage of concepts could provide more comprehensible semantic representations from a user's navigation perspective. To examine the prospects of mixing high and low-level features, the next section explores the impact of weight adjustments on these two promising combinations.

6.3.2 Adjusting feature weights

Following the ratios in the previous section, both mixtures were adjusted by three weight combinations: 1) 80 % keyword weight and 20 % concept weight, 2) 50% keyword weight and 50 % concept weight, and 3) 20 % keyword weight and 80 % concept weight. The second combination (0.5k-0.5c) in Table 7 is exactly the same with the mixture without any weight adjustment (k-c).

Table 7: Means and standard deviations of *Sb* by mixture * weight combination

Mixture	Weight combination	Mean	Std. Deviation	N
0.8k-0.2c	0.8kw+0.2cw.	1.76	.1956	10
	0.5kw+0.5cw	1.53	.1915	10
	0.2kw+0.8cw	1.68	.2027	10
0.5k-0.5c	0.8kw+0.2cw	1.71	.2434	10
	0.5kw+0.5cw	1.56	.1990	10
	0.2kw+0.8cw	1.70	.1519	10

The ANOVA results show that the weight adjustments are significantly different across the mixtures, $p=.011$ (Table 7). The patterns of both mixtures show that weight adjustments do not result in improved map quality. The 50/50 combination without weight adjustment still performs better than any of the other combinations with weight adjustments.

6.4 Feature Combination Analysis in a Four-Book Corpus

Using the single book corpus as discussed earlier, we found that when keywords were combined with concepts, the spread (*Sb*) of the single book corpus was significantly smaller than the one generated only with keywords. When this is repeated using the multiple books corpus, significant differences among various mixture types have to be examined first. The ANOVA shows that there is no significant difference when the keyword/phrase mixtures were used. However, significant differences are found with keyword/concept and keyword/glossary mixtures, $p<.001$ (Table 8).

In addition, there is a significant difference between keyword-only and any other concept combinations in the multiple book corpus, $p=.049$ (Figure 9). The keyword-only results also are significantly different from those with any other glossary combinations in the multiple book corpus, $p<.001$. However, this time, the result is in favor of the keyword approach: the spread for keyword-only maps are of a lower value than those of any other mixtures. A similar pattern is found in the corpus showing that the combination with the higher percentage of keywords can achieve the lowest mean of the spread of the corpus.

Table 8: Means and SDs of *Sb* by corpus*target feature*mixture type in multiple book corpus

Corpus	target feature	mixture type	Mean	Std. Deviation	N
Multiple	phrase	1 k	1.79	.1873	10
		0.8k+0.2p	1.85	.1756	10
		0.5k+0.5p	1.88	.1549	10
		0.2k+0.8p	1.84	.1068	10
		1p	1.89	.1931	10
	concept	1 k	1.79	.1873	10
		0.8k+0.2c	1.90	.1579	10
		0.5k+0.5c	1.90	.1276	10
		0.2k+0.8c	1.94	.2025	10
		1c	2.04	.1940	10
	glossary	1k	1.79	.1873	10
		0.8k+0.2g	1.98	.1117	10
		0.5k+0.2g	2.08	.1935	10

1g	2.74	.1647	10

7. Discussions and conclusions

Researchers have applied SOM to many domains, using keywords as features to represent the content of their corpus and to generate maps. With the increased SOM usage to help users navigate in the information space, an approach to building better-quality SOM is essential. We explored the use of higher-level document representation features to improve the quality of SOM. In addition, we piloted a specific method for evaluating the SOM map quality based on the organization of the information content in the map.

While trying to find more expressive semantic features and to improve the quality of SOM, we examined several features that contained different levels of semantic information and explored their use in building better SOM. Our studies allowed us to discover the following answers to our main research questions:

Q1: Can we produce better SOM by replacing keyword-level document representation with semantic-level representation?

- Keywords are still very powerful content representations in SOM map generation. They outperform any single semantic feature we proposed when measured by the quality of the generated map (although automatically-identified noun phrases produced results which were similar to those with keywords).

Q2: Can we improve the quality of these SOM by enhancing keyword-level document representation with semantic features, and if so, which feature combinations produce the best map?

- Combining keywords with certain semantic features achieves significant improvement in map quality over the keywords-only approach in a relatively homogeneous single book corpus. Changing the ratios in combining different features also affects the performance. Adjusting feature weights does not enhance the performance.

While semantic mixtures can work well in single book corpus, they lose their advantages over keywords in the multiple-book corpus. This raises a concern about whether the semantic representations in the multi-book corpus are homogeneous and coherent enough to apply as semantic features. In a post-analysis study, we found that keyword features presented with the highest coherence rate with 99% of the keywords in the multiple-book corpus also appearing in the single book corpus, while noun phrase and concept features had significantly lower similarity rates, of 82% and 63% respectively. This demonstrates that the terminology issue among textbooks definitely impacts the ability of the SOM to generate a high-quality map for heterogeneous collections. Since the content of a single book has a more consistent semantic representation, the results of the single book study are better than the results of the multiple book study. This once again reinforces the importance of conceptually consistent terms within source content when introducing a semantic approach.

We acknowledge that the lack of positive results of using semantic features in our studies only implies that the specific set of semantic features we have explored are not optimal. There is no implication that semantic representations in general, particularly those high quality-concepts augmented by ontology, are of no use in SOM map construction. In fact, we find that combining semantic features with keywords in the single book corpus offers both tight assemblies of content and improved map quality by providing understandable representations. This shows that semantic features have the potential to enhance the map development.

When employing an alternate evaluation method for SOM quality, our approach of using textbook structure to estimate the content similarity among documents in the corpus was validated. Our study of controlling the various parameters in SOM construction will be useful for the further study of SOM. Our method provides an easy and reasonable evaluation alternative to those domains where the content similarity of documents can be simulated in a similar fashion. Some future research directions are:

- Whether the success of the feature mixture approach that integrated keyword and concept features can be explained by the high recall of relevant documents when using keywords and the high precision when using concepts?
- Whether multiple books by the same author could generate similar results to using the single book corpus?
- Whether better handling of semantic representations, such as using concepts from ontology, could improve the quality of a SOM generated map?

References

AMARASIRI, R., ALAHAKOON, D. & PREMARATHNE, M. Year. The effect of random weight updation in dynamic self organizing maps. *In:* International Conference on Information and Automation (ICIA 2006), 15-17 December 2006 Shandong, China. 183-188.

B RNER, K. & CHEN, C. 2002. Visual interfaces to digital libraries: motivation, utilization, and socio-technical challenges. *Visual Interfaces to Digital Libraries.* Lecture Notes in Computer Science ed. Heidelberg: Springer Berlin.

BAEZA-YATES, R. & RIBEIRO-NETO, B. 1999. *Modern information retrieval,* Boston, MA, Addison Wesley Longman Publishing Co. Inc.

BASILE, P., CAPUTO, A., GENTILE, A. L., DE, M., GEMMIS, LOPS, P. & SEMERARO, G. Year. Improving retrieval experience exploiting semantic representation of documents. *In:* Semantic Web Applications and Perspectives (SWAP 2008), 2008 Rome, Italy.

BELEW, R. K. 2000. *Finding out about,* Cambridge, England, Cambridge University Press.

BENABDESLEM, K. & BENNANI, Y. Year. An incremental SOM for web navigation patterns clustering. *In:* 26th International Conference on Information Technology Interfaces(IEEE 2004), 2004 Cavtat, Italy. 209-213.

BRILL, E. 1993. *A corpus-based approach to language learning.* Doctoral dissertation, University of Pennsylvania.

BRUSILOVSKY, P., CHAVAN, G. & FARZAN, R. Year. Social adaptive navigation support for open corpus electronic textbooks. *In:* SCIENCE, L. N. I. C., ed. 3rd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH'2004), 2004 Eindhoven, the Netherlands. Verlag: Springer, 24-33.

BRUSILOVSKY, P., FARZAN, R. & AHN, J. Year. Comprehensive personalized information access in an educational digital library. *In:* 5th ACM/IEEE-CS Joint Conference on Digital Libraries ACM Press, 2005 Denver, CO. 9-18.

BRUSILOVSKY, P. & RIZZO, R. 2002. Map-based horizontal navigation in educational hypertext. *Journal of Digital Information,* 3**,** 1-10.

CHEN, H., HOUSTON, A. L., SEWELL, R. R. & SCHATZ, B. R. 1998. Internet browsing and searching: user evaluations of category map and concept space techniques. *Journal of the American Society for Information Science,* 49**,** 582-603.

CHEN, H., LALLY, A. M., ZHU, B. & CHAU, M. 2003. HelpfulMed: intelligent searching for medical information over the Internet. *Journal of the American Society for Information Science and Technology,* 54**,** 683-694.

DANG, Y., ZHANG, Y., CHEN, H., HU, P. J., BROWN, S. A. & LARSON, C. 2009. Arizona literature mapper: an integrated approach to monitor and analyze global bioterrorism research literature. *Journal of the American Society for Information Science and Technology,* 60**,** 1466-1485.

DINA, G.-B. & TSVI, K. 2005. Supporting user-subjective categorization with self-organizing maps and learning vector quantization. *Journal of the American Society for Information Science and Technology,* 56**,** 345-355.

ERWIN, E., OBERMAYER, K. & SCHULTEN, K. 1992. Self-organizing maps: ordering, convergence properties and energy functions. *Biological Cybernetics,* 67**,** 47-55.

FARZAN, R. & BRUSILOVSKY, P. Year. Social navigation support through annotation-based group modeling. *In:* INTELLIGENCE, L. N. I. A., ed. 10th International User Modeling Conference, 2005 Edinburgh, UK. Springer Verlag, 463-472.

GONZALO, J., VERDEJO, F., CHUGUR, I. & CIGARRIN, J. Year. Indexing with WordNet synsets can improve text retrieval. *In:* COLING-ACL '98 Workshop on the Usage of WordNet for NLP, 1998 Montreal, Canada. 38-44.

HARPER, M. P. 1992. *The representation of noun phrases in logical form.* Ph.D. Doctoral dissertation, Brown University.

HESKES, T. 1999. Energy functions for Self-Organizing Maps. *Workshop on Self-Organizing Maps (WSOM 99).* Espoo, Finland.

JUN, Y., YOON, H. & CHO, J. 1993. Learning: a fast self-organizing feature map learning algorithm based on Incremental ordering. *IEICE Transactions on Information and Systems,* E76-D**,** 698-706.

KASKI, S. & LAGUS, K. Year. Comparing self-organizing maps. *In:* International Conference on Artificial Neural Networks (ICANN), 1996 Bochum, Germany. London: Springer-Verlag, 809-814.

KIANG, M. Y., KULKARNI, U. R., GOUL, M., PHILIPPAKIS, A., CHI, R. T. & TURBAN, E. Year. Improving the effectiveness of self-organizing map networks using a circular Kohonen layer. *In:* 30th Hawaii International Conference on System Sciences (HICSS) 2006 Maui, HI. 521-530.

KOHONEN, T. 1982. Self-organizing formation of topologically correct feature maps. *Biological Cybernetics,* 43**,** 59-69.

KOHONEN, T. 1990. The self-organizing feature map. *Proceedings of the IEEE,* 78**,** 1464-1480.

KOHONEN, T. 1998. Self-organizing maps. *Neurocomputing,* 21**,** 1-6.

KORFHAGE, R. R. 1997. *Information storage and retrieval,* Hoboken, NJ, John Wiley & Sons, Inc.

LEROY, G. & CHEN, H. 2005. Genescene: an ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts. *Journal of the American Society for Information Science and Technology,* 56**,** 457-468.

LIN, X., SOERGEL, D. & MARCHIONINI, G. Year. A self-organizing semantic map for information retrieval. *In:* 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, 1991 Chicago, IL. ACM, 262-269.

LO, Z.-P. & BAVARIAN, B. 1991. On the rate of convergence in topology preserving neural networks. *Biological Cybernetics,* 65**,** 55-63.

OJA, M., KASKI, S. & KOHONEN, T. 2003. Bibliography of self-organizing map (SOM) papers:1998-2001 addendum. *Neural Computing Surveys,* 3**,** 1-156.

OJA, M., SPERBER, G. O., BLOMBERG, J. & KASKI, S. Year. Grouping and visualizing human endogenous retroviruses by bootstrapping median self-organizing maps. *In:* 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology(CIBCB '04), 2004 La Jolla, CA 95-101.

P LZLBAUER, G. Year. Survey and comparison of quality measures for self-organizing maps. *In:* 5th Workshop on Data Analysis (WDA 2004), 2004 Tatranska Polianka, Slovak Republic. London: Elfa Academic Press, 67-82.

PERUGINI, S., MCDEVITT, K., RICHARDSON, R., P REZ-QUI ONES, M. A., SHEN, R., RAMAKRISHNAN, N., WILLIAMS, C. & FOX, E. A. Year. Enhancing usability in CITIDEL: multimodal, multilingual, and interactive visualization interfaces. *In:* 4th ACM/IEEE-CS joint conference on Digital libraries, 2004 Tuscon, AZ. 315 - 324.

RAUBER, A. & MERKL, D. Year. Using self-organizing maps to organize document archives and to characterize subject matters: how to make a map tell the news of the world. *In:* 10th International Conference on Database and Expert Systems Applications, 1999 Florence, Italy. 302 - 311.

ROUSSINOV, D. G. & CHEN, H. 1998. A scalable self-organizing map algorithm for textual classification: a neural network approach to thesaurus generation. *Communication and Cognition-Artificial Intelligence,* 15**,** 81-111.

SCHATZ, B. R. & CHEN, H. 1996. Introduction to the special issue on building large-scale digital libraries. *IEEE Computer,* 29**,** 22-27.

STOKOE, C., OAKES, M. P. & TAIT, J. Year. Word sense disambiguation in information retrieval revisited. *In:* 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 2003 Toronto, Canada New York: ACM, 159-166.

SU, M.-C., LIU, T.-K. & CHANG, H.-T. 2002. Improving the self-organizing feature map algorithm using an efficient initialization scheme. *Tamkang Journal of Science and Engineering,* 5**,** 35-48.

TANG, H. L. 2002. Knowledge elicitation and semantic representation for the heterogeneous web. *World Wide Web,* 5**,** 229-243.

TEKNOMO, K. *Similarity Measurement.* [Online]. Available: http:\\people.revoledu.com\kardi\ tutorial\Similarity\ [Accessed].

TOLLE, K. M. & CHEN, H. 2000. Comparing noun phrasing techniques for use with medical digital library tools. *Journal of the American Society for Information Science and Technology,* 51**,** 352-370.

TOMURO, N. Year. Question terminology and representation for question type classification. *In:* 19th International Conference on Computational Linguistics (COLING '02), 2002 Taipei, Taiwan. Morristown, NJ: Association for Computational Linguistics, 1-7.

VAN RIJSBERGEN, C. J. 1979. *Information Retrieval,* London, Butterworths.

VICEDO, J. L. & FERR NDEZ, A. Year. A semantic approach to question answering systems. *In:* 9th Text REtrieval Conference, 2000 Gaithersburg, MA. Vdm Verlag, 13-16.

VOUTILAINEN, A. Year. NPtool: a detector of English noun phrases. *In:* Workshop on Very Large Corpora, 1993 Columbus, OH. 48-57.

YANG, C. C., CHEN, H. & HONG, K. 2003. Visualization of large category map for Internet browsing. *Decision Support Systems,* 35**,** 89–102.