Online Information Review

Emerald Article: Improving self-organising information maps as navigational tools: a semantic approach

Yi-ling Lin, Peter Brusilovsky, Daqing He

Yi-ling Lin, Peter Brusilovsky, Daqing He, (2011),"Improving self-organising information maps as navigational tools: a semantic approach", Online Information Review, Vol. 35 Iss: 3 pp. 401 - 424
http://dx.doi.org/10.1108/14684521111151441

Yi-ling Lin, Peter Brusilovsky, Daqing He, (2011),"Improving self-organising information maps as navigational tools: a semantic approach", Online Information Review, Vol. 35 Iss: 3 pp. 401 - 424
http://dx.doi.org/10.1108/14684521111151441

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/1468-4527.htm

Improving self-organising information maps as navigational tools: a semantic approach

Yi-ling Lin, Peter Brusilovsky and Daqing He
School of Information Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Refereed article received
16 February 2010
Approved for publication
5 August 2010

Abstract

Purpose – The goal of the research is to explore whether the use of higher-level semantic features can help us to build better self-organising map (SOM) representation as measured from a human-centred perspective. The authors also explore an automatic evaluation method that utilises 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 – i.e. using only one individual semantic feature, and mixing a semantic feature with keywords. Experiments were conducted to investigate the impact of semantic representation quality on the map. The experiments were performed on data collections from a single book corpus and a 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 in combining different features also affects the performance. While semantic mixtures can work well in a 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 multiple book corpus are homogeneous and coherent enough for applying semantic features. The terminology issue among textbooks affects the ability of the SOM to generate a high quality map for heterogeneous collections.

Originality/value – The authors explored the use of higher-level document representation features for the development of better quality SOM. In addition the authors have piloted a specific method for evaluating the SOM quality based on the organisation of information content in the map.

Keywords Self-organising maps, Semantic representations, Quality evaluation, Feature extraction, Semantics, Maps

Paper type Research paper

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. The 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 for an educational DL (Brusilovsky *et al.*, 2005) found that information maps were the students' most preferred method for accessing information; they were four times more popular

Online Information Review
Vol. 35 No. 3, 2011
pp. 401-424

1468-4527
DOI 10.1108/14684521111151441

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, self-organising maps (SOMs; Kohonen, 1982) are frequently considered to be the most promising mapping approach for large document collections. While they are most popular as a tool for two-dimensional clustering in engineering science, medicine, biology, and economics (Kohonen, 1998; Oja *et al.*, 2003), SOMs are becoming increasingly popular for producing information maps that 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.

However previous studies (Brusilovsky and Rizzo, 2002; Chen *et al.*, 1998) have indicated that the artificial organisation 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 into the same cell, resulting in a negative experience when navigating through SOM. The main challenge of the research presented in this paper was to produce a SOM that provides a closer match to the human conceptualisation of a specific domain. We hypothesise that a potential reason for the "non-human" organisation of SOM is the keyword-level document representation that 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 on 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 resolve the issue 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 to 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 investigates enhancing SOM quality by using semantic features in 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), they did not suggest approaches to evaluate map representation from a human-centred 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 organised as follows: first, a literature review surveys research relevant to self-organising maps. The next section discusses the goal and research questions of this study. The subsequent section introduces our main innovation: the semantic approach to SOM construction. This section also includes the context of our research and the Knowledge Sea information mapping system applied in the study. Then the other innovation, the "textbook" method of SOM evaluation, is proposed. After that we present 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.

Self-organising maps

The self-organising map is a type of an unsupervised neural network model developed by Teuvo Kohonen (1982). A SOM has the ability to reduce the dimensions of data by applying self-organising 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 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 that faithfully preserves the structure of the input data. SOM has spread into numerous fields as a research method, particularly in analysing large volumes of high-dimensional data. The SOM literature can be organised into two branches. One focuses on the study of the relationships between the topical categories. Schatz and Chen's (1996) study showed that SOM has been adopted by many academic projects for textual document classification. Oja *et al.* (2004) categorised human endogenous retroviruses into meaningful groups using SOM. Dina and Tsvi (2005) explored automatic document categorisation methods by comparing SOM and learning vector quantisation. The other branch focuses on an interface for browsing and searching diverse collections. Lin *et al.* (1991) pioneered the use of SOM as a tool for information access. Roussinov and Chen (1998) proposed a multi-level SOM, extending a group of cells into a second layer to assist users with navigation in a large corpus. Rauber and Merkl (1999) showed that the LabelSOM method of automatically labelling the various topical clusters found in the map offered an instant overview for users. Brusilovsky and Rizzo (2002) used 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 to form better maps. Lo and Bavarian (1991) focused on the selection of neighbourhood function, and Kiang *et al.* (2006) proposed a circular training algorithm to overcome the "boundary" effect on topological representations. An incremental learning algorithm had been applied in another study (Jun *et al.*, 1993). Su *et al.* (2002) launched an efficient initialisation scheme to construct an initial map and eventually generate a map with more effective performance.

Studies focusing on SOM representation quality measures are rare in the literature. Most approaches have been mainly concerned with exploring energy functions to

improve the quality of map topology (Erwin *et al.*, 1992; Heskes, 1999). Kaski and Lagus (1996) and Pölzlbauer (2004) compared existing methods for quantifying the quality of SOM. However most existing methods were concerned with topological improvement, not the quality of SOM as a tool for information access. With more SOM applications designed for navigational use, such as multiple layers SOM (Roussinov and Chen, 1998) or an incremental SOM (Benabdeslem and Bennani, 2004), it becomes critical to develop an evaluation method centred on the quality of maps from a human-centred point of view. For most users it is preferable that related contents are grouped together and relationships are easily identifiable, to support their search and browsing. Therefore this paper primarily investigates the ability of SOM to organise content in a similar way to the human approach to content organisation.

Research questions

The goal of the research presented in this paper is to explore whether the use of higher-level semantic features can help us build better SOM representation as measured from a human-centred perspective. Since the higher-level features can be used to produce SOM in two ways (instead 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 these SOM by enhancing keyword-level document representation with semantic features, and if so, which feature combinations produce the best map?

The semantic approach for SOM construction

The problem with building SOM using semantic features can be explored in different contexts. Each context defines a specific combination of a domain and a version of the SOM approach used to organise documents in 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 defines the choice of specific semantic features while the applied map construction approach defines how these features can be used to build 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 the solution. To reflect this importance this section starts with a description of the context that stimulated our research. After that we discuss several semantic features available in the selected domain, and we explain our approach in using these features for SOM construction.

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 Knowledge Sea context 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 have been used in many undergraduate and graduate classes in three domains (Farzan and Brusilovsky, 2005; Brusilovsky *et al.*, 2004; Brusilovsky and Rizzo, 2002):

(1) C-programming;

(2) information retrieval; and

(3) human-computer interaction.

The SOM-based information map in Knowledge Sea is a two-dimensional array of 8×8 cells (Figure 1). Each cell displays a set of keywords and landmarks with features, different icons and background colours. The landmarks provide additional navigation support that helps users locate the cells that contain the most relevant documents. The icons and background provide 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 in the whole map. The properties of SOM ensure that the more similar the documents are, the closer together they are located 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 the documents grouped in these cells. Therefore we could utilise Knowledge Sea as our platform to test SOM quality by evaluating whether relevant textbook documents are assigned to nearby cells 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 for accessing electronic textbooks, outperforming search and browsing (Brusilovsky and Rizzo, 2002). Students also rated 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 this experience that motivated the work presented below.

The domain and the semantic features

The choice of the domain is critical because it defines the kind of semantic features available to be used in SOM construction. For a SOM in a medical domain such as that studied by Chen *et al.* (2003) noun phrases could be the appropriate semantic feature, while 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. In this context the most natural semantic features are domain concepts, which these textbooks are trying to present and explain to the students. These concepts have to be either extracted from the text or provided by experts. For the extraction option we try two state-of-the-art approaches:

(1) noun phrases; and

(2) Yahoo! concepts.

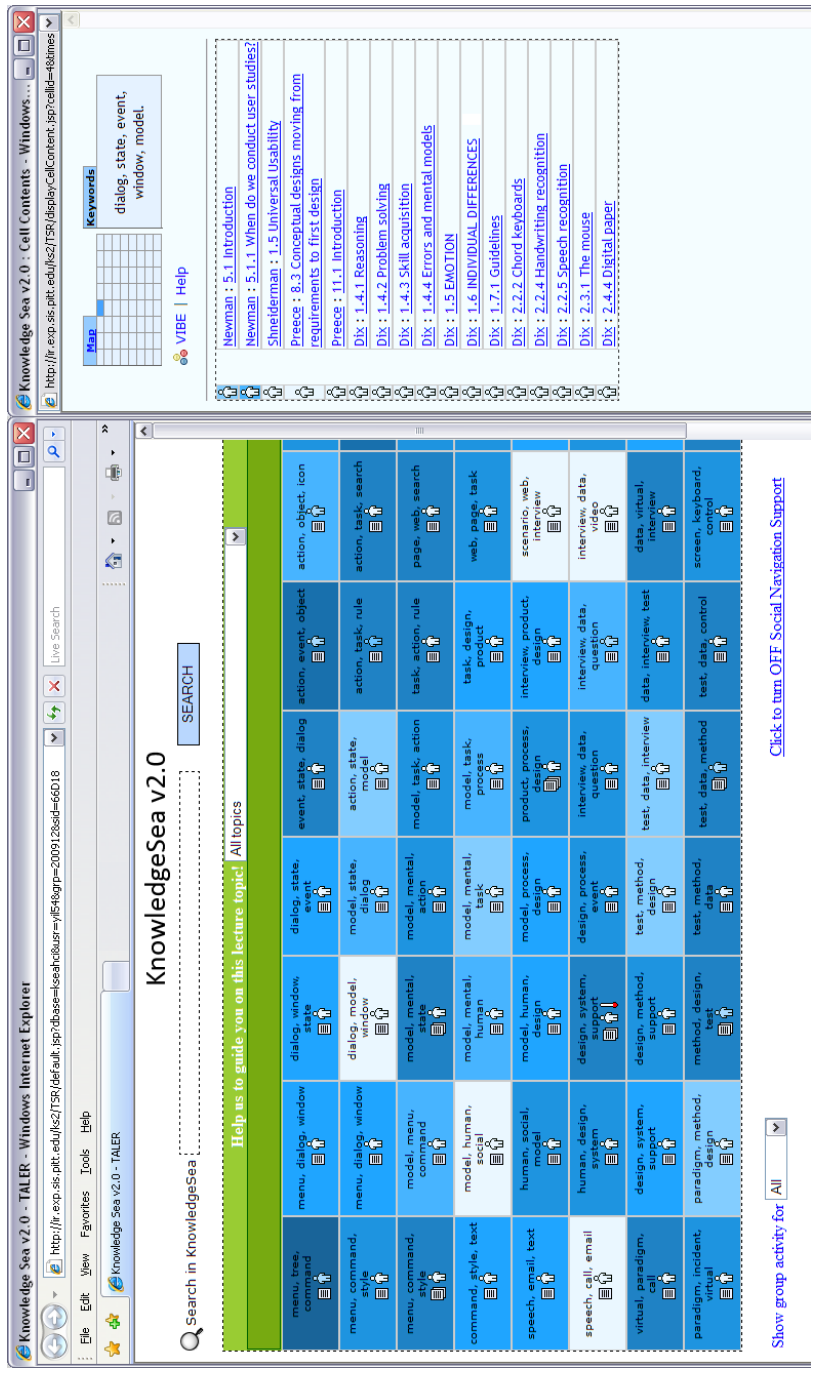

Figure 1.
The system interface of Knowledge Sea

For expert-provided concepts we explore 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 in comparison to traditional keywords.

A keyword is defined as a single term with special significance in the textbook corpus. Standard parsing and tokenisation 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 that is identified by some 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 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 the lowest level semantic features (after keywords). In fact a significant fraction of extracted noun phrases was simply equal to 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 (see http://developer.yahoo.com/search/content/V1/termExtraction.html). The service has been used for a variety of different purposes. For example, Y!Q (http://yq.search.yahoo.com/) uses it to determine key concepts within the search context and apply those concepts to augment a user's search query. Similar to the noun phrases mentioned above, concepts can be considered 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 that was selected by the authors to be included 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 that are dissimilar 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 people are typically selective in picking a set of terms for the glossary.

Our specific interest in exploring glossary terms inspired us to choose a digital library of textbooks on information retrieval (IR) for our study. Among several domains that 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:

(1) *Finding Out About* (Belew, 2000);

(2) *Modern Information Retrieval* (Baeza-Yates and Ribeiro-Neto, 1999);

(3) *Information Retrieval* (Van Rijsbergen, 1979); and

(4) *Information Storage and Retrieval* (Korfhage, 1997).

From a SOM point of view each lowest-level subsection of each textbook is considered a separate document. In total there are 714 documents in the library. The glossary sections of these textbooks contain 402 unique glossary terms.

Semantic map generation

The traditional approach: generating SOM using keyword-level document representation. Using SOM to generate an information map generally involves two steps. The first step involves 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 pre-selected similarity measure such as the cosine similarity or neural network techniques, documents can (one by one) be inserted into the map near or in the most similar cell. The distance between the cells represents the relatedness level among the vectors in these cells. The closer the relationship between vectors of features, the closer the geographic position will be.

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 those of using keywords. Technically the only difference is 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 the 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 weighed vector of semantic features. In total, for each kind of feature, 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 (Figure 2). Generating document representation

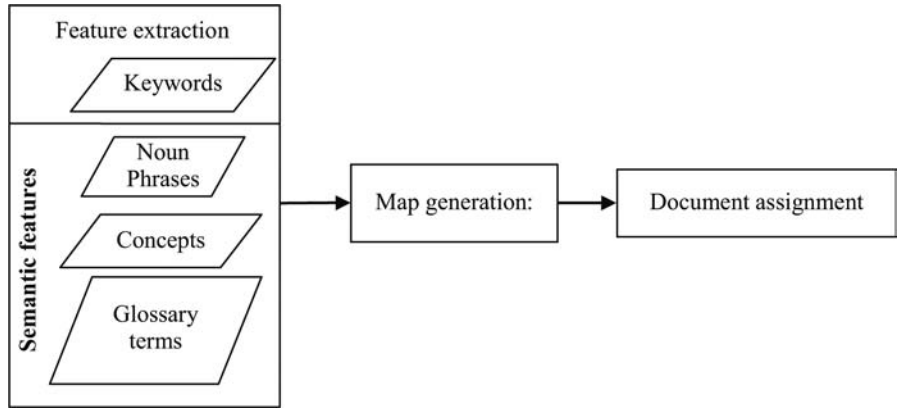

Figure 2.
The SOM generation procedure with semantic features instead of keyword features

and SOMs using mixtures of keywords and various semantic features was done in a more sophisticated way, which is presented in detail in the feature combination analysis section.

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 neighbourhood size. In order to achieve comparable results in our map evaluation we relied on heuristic rules to set these parameters before 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 2,000 times, since any value larger than 2,000 produced almost identical maps and over-representation and under-representation were not issues.

The map size is defined as 8×8 to be consistent with Knowledge Sea. We experimented on three neighbourhood sizes (2, 3, and 4) with ten different learning rates ranging from 0.1 to 0.01. Eventually, through manually checking the generated maps, we defined the neighbourhood 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 is still not determined because SOMs are random by nature. To avoid any side-effects caused by semi-sorted inputs, SOM selects random seeds in the initialisation 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.

The evaluation approach

Evaluating the quality of SOM from a human-centred navigational point of view is a challenging issue that has not been studied thoroughly. As mentioned earlier, existing approaches of SOM evaluation do not take into account the human perspective. A commonly used methodology for user-centred quality evaluation is to apply expert judgments or 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 10 maps for each approach using different random seeds and had to compare two sets of maps, rather than simply comparing two 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 participant 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 structure. Moreover it is not simply a random user judgment (as we would get in 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.

A textbook-centred evaluation approach

The textbook-centred evaluation approach that we propose is based on the properties of academic textbooks. By their design textbooks focus on a specific issue (a topic) of the domain in each chapter. Within a chapter (first level), more specific concepts related to the chapter's key issue are systematically examined section (second level) by section, 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, because they are 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 organisation of the domain. In this study we defined a cluster as a section in a chapter in a textbook. The assumption is that a more human-centred SOM construction approach, the one that better preserves the conceptual structure of the domain 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 3 Section 1-1-1 is conceptually close to Section 1-1-2, Section 1-1-3, and Section 1-1-4 but quite distant from Section 4-3-1 (section 1-2-3 means chapter 1, section 2, subsection 3). Thus a good map should display Section 1-1-1 and Section 1-1-2 closer together than Section 1-1-1 and Section 4-3-1. Figure 4 places Section 1-1-1 closer to Section 4-3-1 than to Section 1-1-2 or Section 1-1-3, which may indicate a conceptual problem with this map.

In this study we assessed map quality by calculating average corpus spread in three steps:

(1) The spread of two documents is defined by the Euclidean distance (Teknomo, 2006) between the cells that documents D1 and D2 are in (X and Y):

$$Lp(D1, D2) = \sqrt{\sum\nolimits_I (d1i - D2i)^2},\ (i = X \text{ dimension},\ Y \text{ dimension}). \quad (1)$$

For example, if Section 1-1-1 is located in cell (0,0) and Section 1-1-2 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:

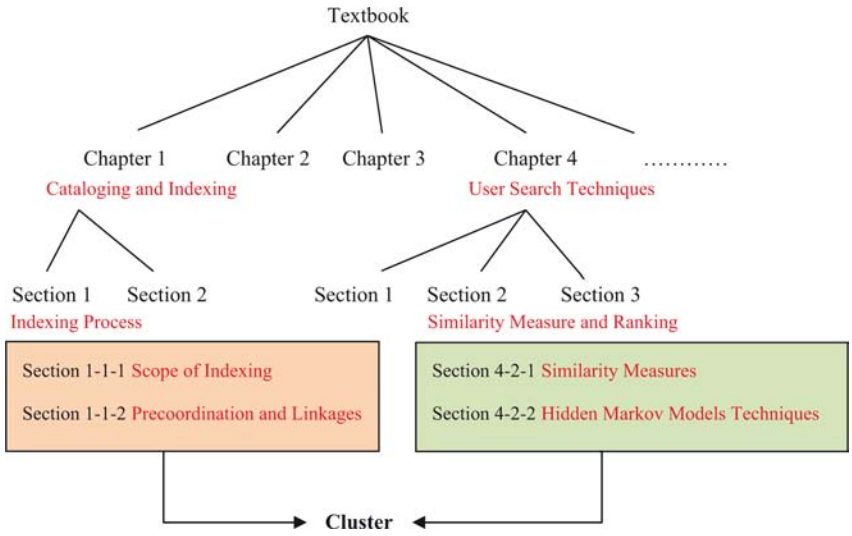

Figure 3.
Textbook structure

Section 1-1-1		
	Section 4-3-1	Section 1-1-3
Section 1-1-2		Section 4-3-2

Figure 4.
The map with incorrect document assignments

$$\mathrm{Sc} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{LP(Di,Dj)}{2}\right]}{n!/(n-2)!2!} \ (n \text{ is the number of documents in the cluster}). \quad (2)$$

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

$$\mathrm{Sb} = \frac{\sum_{i=1}^{n}\mathrm{Sc}_1}{n} \ (n \text{ is the number of clusters}). \quad (3)$$

Feasibility examination

Our evaluation approach is based on the assumption that documents within a textbook cluster (i.e. subsections of the same section) are more similar to each other than to documents outside 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 I 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 subsections from different clusters. The Wilcoxon signed ranks test shows that this difference is significant ($p < 0.001$) 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 that cluster.

The study

Motivated by the research questions presented earlier, we conducted a set of experiments to investigate the impact of semantic representations on 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 a significant improvement in map quality over the keywords-only approach.

The experiments were performed on data collections from one 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 semantic feature; and

(2) mixing a semantic feature with keywords.

Therefore the experiments were modelled as several ANOVA experiments. The dependent variable is the spread of the corpus, 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 ten 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 ten maps generated using the feature.

Individual feature analysis in a four-book corpus

In this stage we used all four books to generate maps. We were interested in comparing the spread for the four individual representations (i.e. how far 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 represents the total volume of the glossary collection.

Table I. Cross-cluster and within-cluster average similarity

Book	Cross-cluster	Within-cluster
Belew (2000)	0.10	0.28
Baeza-Yates and Ribeiro-Neto (1999)	0.08	0.35
Van Rijsbergen (1979)	0.07	0.38
Korfhage (1997)	0.14	0.40

As shown in Table II, contradictory 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 for keywords with ten random seeds has the lowest value (1.79) and also produces the minimum value (1.55) among all 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 keywords (Table II).

The ANOVA results show that there is a significant difference among the features at $p < 0.001$. The mean of the spread for keywords with ten random seeds is significantly lower than that for concepts, $p = 0.002$, and glossary terms, $p < 0.001$ (Figure 5). 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 analysing and

	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

Note: R1-R10 indicate ten different random seeds

Table II.
The spread of the multiple book corpus with a constant set of ten random seeds

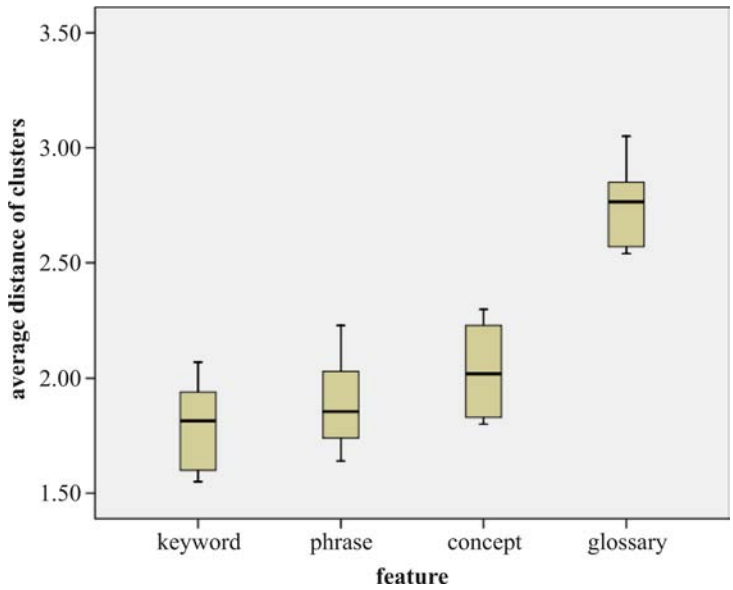

Figure 5.
The boxplot of features in the multiple book corpus

merging the glossaries from these books. These glossaries are substantially different from each other, with almost no overlap. Out of 402 glossary terms extracted from the four glossaries, only nine terms appear in more than two books and no terms appear in three or more (Table III).

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 III 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.

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.

The results show that the mean of the spread for phrases with ten random seeds has the lowest value and also produces the minimum value among all results (Table IV).

Table III.
Indexing heterogeneity for different features

Number of features shared by …	Keywords	Noun phrases	Concepts	Glossary terms
One book	0	47	307	393
Two books	9	100	117	9
Three books	88	154	114	0
All four books	503	299	62	0
Total	600	600	600	402

However according to the ANOVA results, the mean of the spread for phrase features is not significantly different from the one for keyword features, $p = 0.997$. 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 outperformed keywords, although phrases performed equally well. The performances of concept ($p = 0.022$) and glossary ($p < 0.001$) are still significantly worse than the performance of keywords (Figure 6).

The poor performance of concepts and glossary items in a single book corpus demonstrate that heterogeneity may not be the most critical difference between indexing with higher-level features and with traditional keywords. To study the problem more deeply we compared low-level differences between several kinds of indexing. The most interesting issue is indexing density: how many features of different levels can be found on a single page and, vice versa, how many pages are indexed with 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 indexing density fell rapidly (Table V). In keyword-level indexing, each book section is represented by 600 high-frequency keywords, with 77.88 unique keywords per page and almost 200 recognised keywords

	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

Note: R1-R10 indicate ten different random seeds

Table IV.
The spread of the single book corpus with a constant set of 10 random seeds

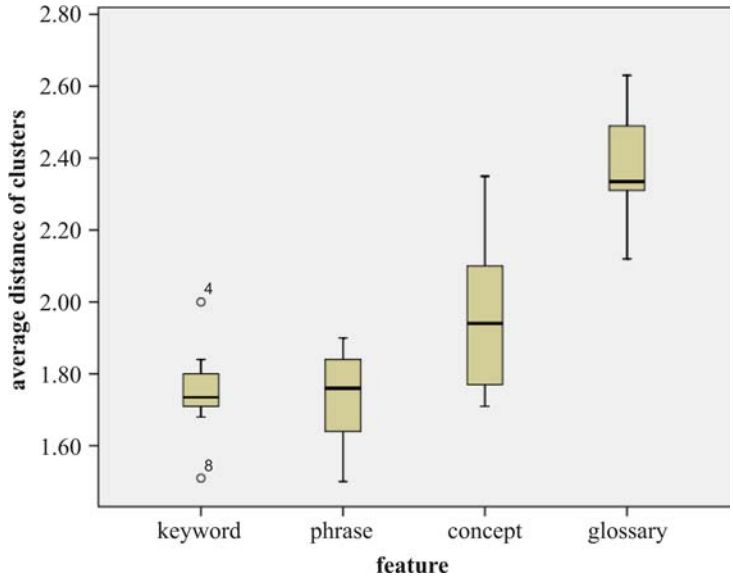

Figure 6.
The boxplot of features in single book corpus

overall. On the other end of the spectrum each section is represented on average by only 6.13 unique glossary terms. Noun phrases rank very close to keywords (most of them being, in fact, single nouns) while concepts stand somewhat between the two extremes. The low density of indexing shows clearly 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 points out that this representation may be "patchy"; some aspects of the page content are not represented at all. This fact is additionally confirmed by the significant increase in the number of pages that have none of the features listed in the top 600 concepts or the top 402 glossary items (Table V). It is interesting to observe that the performance of higher-level features (Table IV) does not decrease as rapidly as indexing density (Table V). 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 cannot compensate for the rapid fall of indexing density and the resulting patchy representation of units.

One possible way to increase the density of indexing while maintaining the semantic depth of representation could be a radical increase in the number of features used for indexing (i.e. from 600 top features to 2,000 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 of them. 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 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 units' 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 potential.

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 that of our 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 among the features, and the other is focused on adjusting the weights of the features.

Table V.
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

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 the single book corpus. Keeping the total number of features constant we explored five different ratio combinations:

(1) keyword only;

(2) 80 per cent keyword and 20 per cent target feature;

(3) 50 per cent keyword and 50 per cent target feature;

(4) 20 per cent keyword and 80 per cent target feature; and

(5) target feature only.

For example, the keyword-only combination had 600 keywords, whereas the 80 per cent keyword and 20 per cent target feature combination had 480 keywords and 120 target feature terms (Table VI). The whole process of ten map generation, section assignment, and distance calculations was performed for each of these combinations. The ANOVA results show that the mixture of keyword and phrases is not able to outperform keywords significantly.

As Table VI shows, the use of feature mixtures does affect the quality of resulting SOM. For each of the three higher-level features there is at least one combination that produces better results than single keywords alone. Most importantly, we found a significant difference between keyword-only and any other keyword/concept combination in the single book corpus, $p = 0.005$ (Figure 7). In fact, any keyword/concept combination performed better than keywords alone. In addition we

Corpus	Target feature	Mixture type	Mean	SD
Single	Phrase	1k	1.75	0.1248
		0.8k + 0.2p	1.67	0.1686
		0.5k + 0.5p	1.78	0.2136
		0.2k + 0.8p	1.76	0.1562
		1p	1.74	0.1283
	Concept	1k	1.75	0.1248
		0.8k + 0.2c	1.53**	0.1915
		0.5k + 0.5c	1.56**	0.1990
		0.2k + 0.8c	1.65**	0.1413
		1c	1.96	0.1998
	Glossary	1k	1.75	0.1248
		0.8k + 0.2g	1.70	0.1559
		1g	2.38	0.1423

Notes: All $n = 10$ (maps). **Significant at $p < 0.01$

Table VI.
Means and SDs of Sb by corpus*target feature*mixture type in the single book corpus

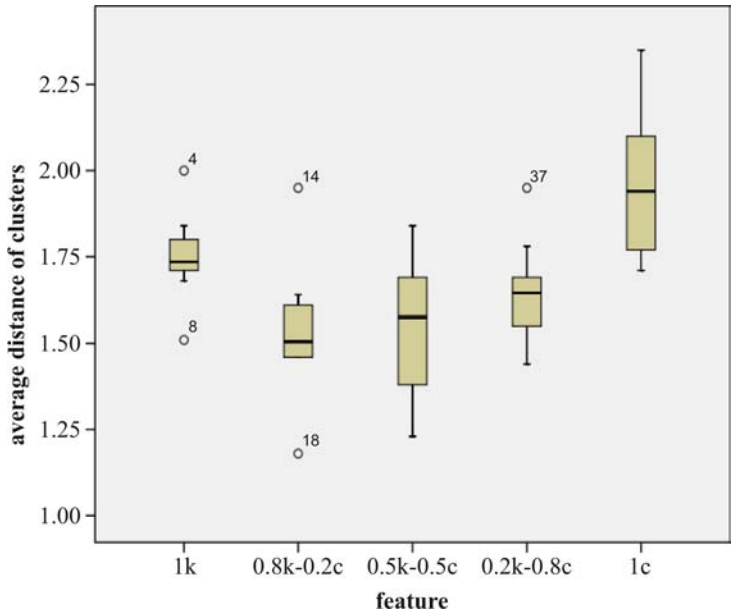

Figure 7.
The boxplot of the mixture of keywords and concepts in the single book corpus

observed that slightly better results are achieved when the keyword ratio is set as the higher of the two ratios in the combination.

Next the study moved on to compare keyword-only with each combination, and to look for the best combination of keyword and concept mixtures in the single book corpus. The marginal comparisons reveal that the keyword-only approach has a significantly larger mean of the spread than the combination of 80 percent keyword and 20 per cent concept, $p = 0.004$, or the combination of 50 per cent keyword and 50 per cent concept, $p = 0.009$ (Table VI). Even though the combination of 80 per cent keyword and 20 per cent concept has a lower mean of the spread, which means better performance than the combination of 50 per cent keyword and 50 per cent 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.

Adjusting feature weights. Following the ratios in the previous section both mixtures were adjusted by three weight combinations:

(1) 80 per cent keyword weight and 20 per cent concept weight;

(2) 50 per cent keyword weight and 50 per cent concept weight; and

(3) 20 per cent keyword weight and 80 per cent concept weight.

The second combination below is exactly the same with the mixture without any weight adjustment.

The ANOVA results show that the weight adjustments are significantly different across the mixtures, $p = 0.011$ (Table VII). The patterns of both mixtures show that

weight adjustments do not improve map quality. The 50/50 combination without weight adjustment still performs better than other combinations with weight adjustments.

Feature combination analysis in a four book corpus

When we used the single book corpus, as reported above, we found that when keywords were combined with concepts the spread of the single book corpus was significantly smaller than the one generated with only keywords. When this is repeated using the multiple book corpus, significant differences among different mixture types have to be examined first. The ANOVA shows that there is no significant difference with the keyword/phrase mixtures. However significant differences are found with keyword/concept and keyword/glossary mixtures, $p < 0.001$ (Table VIII).

In addition there is a significant difference between keyword-only and any other concept combination in the multiple book corpus, $p = 0.049$ (Figure 8). The keyword-only results also have a significant difference with any other glossary combination in the multiple book corpus, $p < 0.001$. However this time the result is in favour of the keyword approach: the spread for keyword-only maps is of a lower value than any of the other mixtures. A similar pattern is found in the corpus, showing that

Mixture	Weight combination	Mean	SD
0.8k-0.2c	0.8kw + 0.2cw	1.76	0.1956
	0.5kw + 0.5cw	1.53	0.1915
	0.2kw + 0.8cw	1.68	0.2027
0.5k-0.5c	0.8kw + 0.2cw	1.71	0.2434
	0.5kw + 0.5cw	1.56	0.1990
	0.2kw + 0.8cw	1.70	0.1519

Note: All $n = 10$

Table VII. Means and SDs of Sb by mixture*weight combination

Corpus	Target feature	Mixture type	Mean	SD
Multiple	Phrase	1k	1.79	0.1873
		0.8k + 0.2p	1.85	0.1756
		0.5k + 0.5p	1.88	0.1549
		0.2k + 0.8p	1.84	0.1068
		1p	1.89	0.1931
	Concept	1k	1.79	0.1873
		0.8k + 0.2c	1.90	0.1579
		0.5k + 0.5c	1.90	0.1276
		0.2k + 0.8c	1.94	0.2025
		1c	2.04	0.1940
	Glossary	1k	1.79	0.1873
		0.8k + 0.2g	1.98	0.1117
		0.5k + 0.2g	2.08	0.1935
		1g	2.74	0.1647

Note: All $n = 10$

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

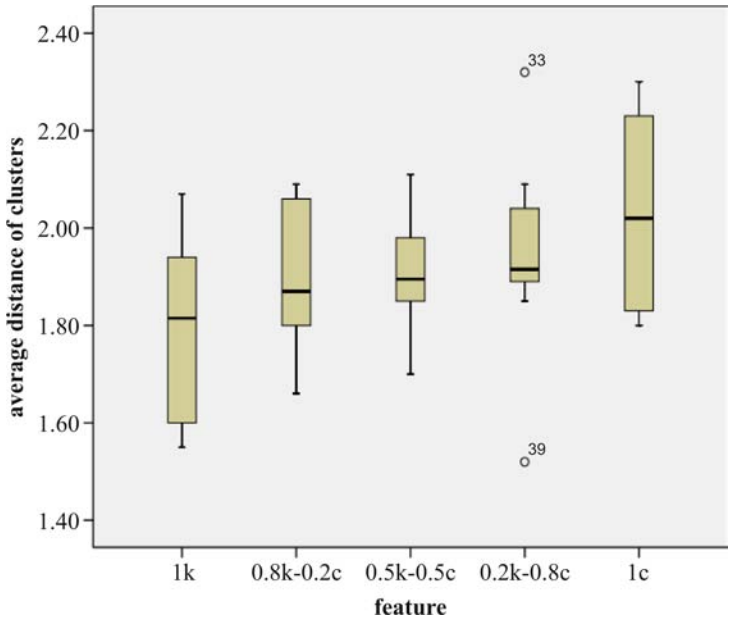

Figure 8.
The boxplot of the mixture of keyword and concept in the multiple book corpus

the combination with the higher percentage of keywords can achieve the lowest mean of the spread of the corpus.

Discussions and conclusions

Researchers have applied SOM in many domains using keywords as features to represent the content of their corpus and generate maps. With the increased usage of SOM to help users navigate in the information space, an approach to build better quality SOM is required. 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 map quality based on the organisation of information content in the map.

On the way to finding 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 for building better SOM. Our studies allowed us to give 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 generated map quality (although automatically identified noun phrases produced results which were not significantly different from 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 of 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 a 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 multiple book corpus are homogeneous and coherent enough for applying semantic features. In a post-analysis study we found that keyword features showed the highest coherence rate with 99 per cent 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, with 82 per cent and 63 per cent, respectively. This demonstrates that the terminology issue among textbooks definitely affects the ability of the SOM to generate a high quality map for heterogeneous collections. Since single book content 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 from using semantic features in our studies only implies that the 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 with ontology, are of no use in SOM map construction. In fact we found that combining semantic features with keywords in the single book corpus not only has tight assemblies of content but also improves map quality by providing understandable representations. This shows that semantic features have the potential to enhance map usage.

In the alternative method for evaluating 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 for the domains whose documents' content similarity can be simulated in similar fashion. Some future directions are:

- Whether the success of the feature mixture approach that integrated keyword and concept features can be explained by keyword's high recall of relevant documents and concept's high precision to users' requirements?
- Whether multiple books by the same author could generate similar results 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. and Premarathne, M. (2006), "The effect of random weight updation in dynamic self organizing maps", *Proceedings of the International Conference on Information and Automation*, IEEE, New York, NY, pp. 183-8.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, Addison-Wesley Longman, Boston, MA.

Basile, P., Caputo, A., Gentile, A.L., De, M., Lops, P. and Semeraro, G. (2008), "Improving retrieval experience exploiting semantic representation of documents", paper presented at Semantic Web Applications and Perspectives, Rome, 15-17 December.

Belew, R.K. (2000), *Finding Out About*, Cambridge University Press, Cambridge.

Benabdeslem, K. and Bennani, Y. (2004), "An incremental SOM for web navigation patterns clustering", *Proceedings of the 26th International Conference on Information Technology Interfaces*, IEEE, New York, NY, pp. 209-13.

Börner, K. and Chen, C. (2002), "Visual interfaces to digital libraries: motivation, utilization, and socio-technical challenges", *Visual Interfaces to Digital Libraries*, Lecture Notes in Computer Science, Vol. 2539, Springer, Berlin, pp. 1-9.

Brill, E. (1993), "A corpus-based approach to language learning", PhD thesis, University of Pennsylvania, Philadelphia, PA.

Brusilovsky, P. and Rizzo, R. (2002), "Map-based horizontal navigation in educational hypertext", *Journal of Digital Information*, Vol. 3 No. 1, pp. 1-10.

Brusilovsky, P., Chavan, G. and Farzan, R. (2004), "Social adaptive navigation support for open corpus electronic textbooks", *Proceedings of the 3rd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, Springer, Berlin, pp. 24-33.

Brusilovsky, P., Farzan, R. and Ahn, J. (2005), "Comprehensive personalized information access in an educational digital library", *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM Press, New York, NY, pp. 9-18.

Chen, H., Houston, A.L., Sewell, R.R. and 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*, Vol. 49 No. 7, pp. 582-603.

Chen, H., Lally, A.M., Zhu, B. and Chau, M. (2003), "HelpfulMed: intelligent searching for medical information over the internet", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 7, pp. 683-94.

Dang, Y., Zhang, Y., Chen, H., Hu, P.J., Brown, S.A. and 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*, Vol. 60 No. 7, pp. 1466-85.

Dina, G.-B. and 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*, Vol. 56 No. 4, pp. 345-55.

Erwin, E., Obermayer, K. and Schulten, K. (1992), "Self-organizing maps: ordering, convergence properties and energy functions", *Biological Cybernetics*, Vol. 67 No. 1, pp. 47-55.

Farzan, R. and Brusilovsky, P. (2005), "Social navigation support through annotation-based group modelling", *Proceedings of the 10th International User Modeling Conference, Edinburgh*, Lecture Notes in Artificial Intelligence, Springer, Berlin, pp. 463-72.

Gonzalo, J., Verdejo, F., Chugur, I. and Cigarrin, J. (1998), "Indexing with WordNet synsets can improve text retrieval", *COLING-ACL '98 Workshop on the Usage of WordNet for NLP, Montreal*, pp. 38-44.

Harper, M.P. (1992), "The representation of noun phrases in logical form", PhD thesis, Brown University, Providence, RI.

Heskes, T. (1999), "Energy functions for self-organizing maps", *Proceedings of the Workshop on Self-Organizing Maps*, Springer, Berlin, pp. 301-15.

Jun, Y., Yoon, H. and Cho, J. (1993), "Learning: a fast self-organizing feature map learning algorithm based on incremental ordering", *IEICE Transactions on Information and Systems*, Vol. E76, pp. 698-706.

Kaski, S. and Lagus, K. (1996), "Comparing self-organizing maps", in *Proceedings of the International Conference on Artificial Neural Networks*, Springer, London, pp. 809-14.

Kiang, M.Y., Kulkarni, U.R., Goul, M., Philippakis, A., Chi, R.T. and Turban, E. (2006), "Improving the effectiveness of self-organizing map networks using a circular Kohonen layer", *Proceedings of the 30th Hawaii International Conference on System Sciences*, IEEE Computer Society, Washington, DC, pp. 521-30.

Kohonen, T. (1982), "Self-organizing formation of topologically correct feature maps", *Biological Cybernetics*, Vol. 43 No. 1, pp. 59-69.

Kohonen, T. (1990), "The self-organizing feature map", *Proceedings of the IEEE*, Vol. 78 No. 9, pp. 1464-80.

Kohonen, T. (1998), "Self-organizing maps", *Neurocomputing*, Vol. 21 Nos 1-3, pp. 1-6.

Korfhage, R.R. (1997), *Information Storage and Retrieval*, Wiley, Hoboken, NJ.

Leroy, G. and 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*, Vol. 56 No. 5, pp. 457-68.

Lin, X., Soergel, D. and Marchionini, G. (1991), "A self-organizing semantic map for information retrieval", *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, NY, pp. 262-9.

Lo, Z.-P. and Bavarian, B. (1991), "On the rate of convergence in topology preserving neural networks", *Biological Cybernetics*, Vol. 65 No. 1, pp. 55-63.

Oja, M., Kaski, S. and Kohonen, T. (2003), "Bibliography of self-organizing map (SOM) papers: 1998-2001 addendum", *Neural Computing Surveys*, Vol. 3 No. 1, pp. 1-156.

Oja, M., Sperber, G.O., Blomberg, J. and Kaski, S. (2004), "Grouping and visualizing human endogenous retroviruses by bootstrapping median self-organizing maps", *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, IEEE Press, Los Alamitos, CA, pp. 95-101.

Perugini, S., McDevitt, K., Richardson, R., Pérez-Quiñones, M.A., Shen, R., Ramakrishnan, N., Williams, C. and Fox, E.A. (2004), "Enhancing usability in CITIDEL: multimodal, multilingual, and interactive visualization interfaces", *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM Press, New York, NY, pp. 315-24.

Pölzlbauer, G. (2004), "Survey and comparison of quality measures for self-organizing maps", *Proceedings of the 5th Workshop on Data Analysis*, Elfa Academic Press, London, pp. 67-82.

Rauber, A. and Merkl, D. (1999), "Using self-organizing maps to organize document archives and to characterize subject matters: how to make a map tell the news of the world", *Proceedings of the 10th International Conference on Database and Expert Systems Applications*, Springer, Berlin, pp. 302-11.

Roussinov, D.G. and Chen, H. (1998), "A scalable self-organizing map algorithm for textual classification: a neural network approach to thesaurus generation", *Communication and Cognition-Artificial Intelligence*, Vol. 15 Nos 1/2, pp. 81-111.

Schatz, B.R. and Chen, H. (1996), "Introduction to the special issue on building large-scale digital libraries", *IEEE Computer*, Vol. 29 No. 5, pp. 22-7.

Stokoe, C., Oakes, M.P. and Tait, J. (2003), "Word sense disambiguation in information retrieval revisited", *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, NY, pp. 159-66.

Su, M.-C., Liu, T.-K. and Chang, H.-T. (2002), "Improving the self-organizing feature map algorithm using an efficient initialization scheme", *Tamkang Journal of Science and Engineering*, Vol. 5 No. 1, pp. 35-48.

Tang, H.L. (2002), "Knowledge elicitation and semantic representation for the heterogeneous web", *World Wide Web*, Vol. 5 No. 3, pp. 229-43.

Teknomo, K. (2006), "Similarity measurement", available at: http://people.revoledu.com/kardi/tutorial/Similarity/ (accessed December 10, 2009).

Tolle, K.M. and Chen, H. (2000), "Comparing noun phrasing techniques for use with medical digital library tools", *Journal of the American Society for Information Science and Technology*, Vol. 51 No. 4, pp. 352-70.

Tomuro, N. (2002), "Question terminology and representation for question type classification", *Proceedings of the 19th International Conference on Computational Linguistics*, Association for Computational Linguistics, Morristown, NJ, pp. 1-7.

Van Rijsbergen, C.J. (1979), *Information Retrieval*, Butterworths, London.

Vicedo, J.L. and Ferrández, A. (2000), "A semantic approach to question answering systems", *Proceedings of the 9th Text Retrieval Conference, Gaithersburg, MA*, VDM Verlag, Saarbrücken, pp. 13-16.

Voutilainen, A. (1993), "NPtool: a detector of English noun phrases", *Proceedings of the Workshop on Very Large Corpora, Columbus, OH*, pp. 48-57.

Yang, C.C., Chen, H. and Hong, K. (2003), "Visualization of large category map for internet browsing", *Decision Support Systems*, Vol. 35 No. 1, pp. 89-102.

About the authors

Yi-ling Lin is a PhD student in Information Sciences at the University of Pittsburgh and works in the Personalized Adaptive Web Systems Lab on educational and cultural heritage projects. She has been working in the field of human interactive systems, information retrieval and the social web for three years. She has published papers about semantic enhancement and human interactive systems in journals and conference proceedings and also participates actively in academic services, such as reviewing papers and volunteering. This project cooperated with the Carnegie Museum of Art and an internship at the Free University of Amsterdam in The Netherlands by joining the NWO-CATCH-2 Project. Her research involves investigating social wisdom in classification and indexing tasks. Yi-ling Lin is the corresponding author and can be contacted at: yil54@pitt.edu

Peter Brusilovsky has been working in the field of adaptive educational systems, user modeling, and social information systems for more than 20 years. He is currently an Associate Professor of Information Science and Intelligent Systems at the University of Pittsburgh, where he directs the Personalized Adaptive Web Systems Lab. Peter is the Associate Editor-in-Chief of *IEEE Transactions on Learning Technologies* and a board member of several journals, including *User Modeling and User Adapted Interaction*, *ACM Transactions on the Web*, and *Web Intelligence and Agent Systems*. He is also the current President of User Modeling, Inc., a professional association of user modeling researchers.

Daqing He is an Associate Professor in the School of Information Sciences at the University of Pittsburgh. His research interests focus on information retrieval, intelligent information systems, and digital libraries.