Iterative Visual Clustering for Unstructured Text Mining

Qian You, Shiaofen Fang
Dept of Computer and Information Science,
Indiana-University Purdue-University
723 W. Michigan Street SL280, Indianapolis, IN46202
+1-317-278-2293
{qiyou, sfang}@cs.iupui.edu

Patricia Ebright
Dept. of Adult Health,
Indiana University
1111 Middle Dr., MU442, Indianapolis, IN46202
+1-317-274-7942
prebrigh@iupui.edu

ABSTRACT

This paper proposes the iterative visual clustering (IVC) on unstructured text sequences to form and evaluate keyword clusters, based on which users can use visual analysis, domain knowledge to discover knowledge in the text. The text sequence data are broken down into a list representative keywords after textual evaluation, and the keywords are then grouped to form keyword clusters via an iterative stochastic process and are visualized as distributions over the time lines. The visual evaluation model provides shape evaluations as quantitative tools and users' interactions as qualitative tools to visually investigate the trends, patterns represented by the keyword clusters' distributions. The keyword clustering model, guided by the feedback of visual evaluations, step-wisely enumerates newer generations of keyword clusters and their patterns, therefore narrows down the search space. Then the proposed IVC is applied onto nursing narratives and is able to identify interesting keyword clusters implying hidden knowledge regarding to the working patterns and environment of registered nurses. The loop of producing next generation of keyword clusters in IVC is driven and controlled by users' perception, domain knowledge and interactions, and it is also guided by a stochastic search model. So both semantic and distribution features enable IVC to have significant applications as a text mining tool, on many other data sets , such as biomedical literatures.

Categories and Subject Descriptors

I.3.5 [**Computer Graphics**]: Computational Geometry and Object Modeling-Geometric algorithms. J.3 [**Computer Applications**]: Life and Medical Science; Health, Medical Information Systems.

General Terms

Algorithms, Design, Human Factors

Keywords

Text and document visualization, Nursing data processing.

1.INTRODUCTION

Recently human intuitions and domain expertise are integrated into the learning process of data mining applications, to assist knowledge discovery. The integration requires a more dynamic and exploratory visualization environment that provides efficient visual representations for the users to analyze the profile of the problem, to translate users' interactions as input and constraints to the data mining flow and then to represent analysis results as interpretable visual information.

Unstructured texts, such as real time narratives and daily notes, are one of the major sources for data mining, because of its abundance and easy access. However, Natural language parsing (NLP) will be less effective on unstructured text as the language structures are unreliable or non-existing. So it is critical to have intuitive and interactive methods to navigate and explore unstructured text data to perceive, analyze and discover potential patterns, trends or other hidden information, in the vast quantities of unstructured information. The problem is particularly important for health care applications, as most clinical and patient care records and narratives are unstructured text sequences. For example, knowledge inferred from keyword clusters of nursing narratives is the key to understand the nursing working environment and tasks which can further help to redesign the nurses' assignments and to improve nurses' daily performance.

The proposed iterative visual clustering (IVC) is motivated by the need of integrating users' inputs into the system learning loop, and then uses visualization techniques to interface user interactions and the system. In IVC, users' judgments on keywords clusters' meaning and their pattern shapes, together with a global distance function, plays the role of objective functions. Both automatic and interactive evaluations are coded to change the probability distributions in a stochastic search process, which in turn produces the next generation of keywords clusters. Yet the number of all possible combinations of keywords derived from the text data is huge, so a brute force way to sample every combination of keywords is infeasible. To properly address the problems of searching keyword clusters, evaluating keyword clusters and generating successor clusters, IVC is broken into several steps and the diagram of our process is shown in Figure 1. First we start with a list of representative keywords, each of them being a single-word cluster and being visualized as a distribution. Second, the **automatic evaluation** linearly ranks the thumbnails

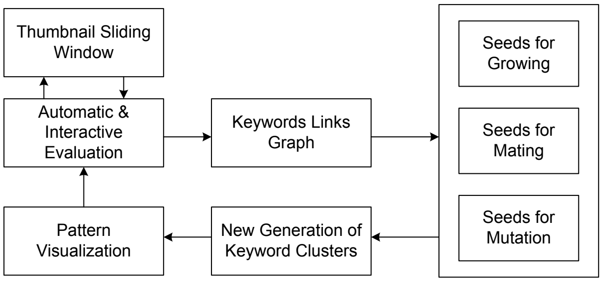


Figure 1. Diagram and data flow of iterative visual clustering (IVC)

ISB'10, February 15-17, 2010, Calicut, India.

of cluster patterns with a global distance function and visualizes the dispersion of the shapes of keyword clusters. Third, **interactive evaluation,** working with automatic evaluation, filters out unsatisfied keyword clusters, updates the feedback and stochastically produces three sets of clusters including **seeds for growing, seeds for mating and seeds for mutation.** Finally, clusters in the aforementioned three sets evolve into a new generation of cluster population accordingly. In the loop of IVC, users domain knowledge and interactions play critical roles in every step: they evaluate the shapes of the clusters, identify seeds need to be grown and control the dispersion and size of keywords clusters. We also apply our proposed iterative visual clustering (IVC) method onto nursing narratives to discovery important keyword clusters inferring the knowledge of working environment.

In summary IVC have contributions in the following aspects:

1) IVC helps users detect interesting keyword clusters and hidden knowledge in unstructured text by visualizing and organizing a large number of keyword clusters and their distribution patterns in an iterative searching process.
2) Users help IVC narrow down the search space by evaluations that combine visualization, geometry processing, and improve the search process by interactively updating the probability of the stochastic search direction.
3) Applying IVC onto nursing narratives, meaningful keywords clusters and their implied knowledge are discovered, leading to understandings and possible redesign of nurse working patterns and working environment.

In the following context, we first review previous work, then describe the procedures of text pre-processing and introduce the visualization of keyword clusters. After that, the visual evaluation model for keyword cluster patterns is formalized and the users' interactive evaluations are described, followed by the elaborations of the stochastic model of generating new clusters. Then the results of applying IVC onto nursing data are present and discussed. Finally, we conclude the paper with additional remarks and future work.

2.RELATED WORK

Over the past decades, there has been much research dedicating to document visualization, and they usually present the overview of document collections [3][5] which are then clustered by unsupervised learning algorithms [13]. Other text and document visualization systems [1][21] investigate the keyword clusters and classes within the free-form running text by mapping text mining results onto visual metaphors, among which are images [11], concept map [2], Implicit Surface [18]. Some of the visualization techniques used in health information data are related to IVC as they visualize the text depending on Keywords Distribution [6]. CareView [12], for example, visualizes integrated trends of both quantitative and qualitative data in clinical records and narratives. LifeLines [17]describes a visualization environment which provides an overview of a series of events such as personal histories, medical records, and court records. ThemeRiver [9] uses the natural image of a flowing river as a visual metaphor to illustrate changes of themes over time. While these works certainly introduce 'visual information' to understand the non-visual text, the visualization processes are separated from the text mining and analysis step, and few users' feedback are taken in to drive the evolving of the system.

To feedback users' judgment on the 'goodness' of visual representations, researchers in Computer Graphics have developed methodologies, such as Design Gallery [14], to explore the parameter space of transfer functions by visualizing the most dispersed image populations. Users pick up the best images by observing them hence simplify the trial-error process of parameter tweaking. The visual system in [20] also visualizes the patterns of a great many keywords and forms the clusters of keywords purely based on the shapes of their patterns. However neither the quantitative shape evaluations nor the cluster generation process is formalized. We found that visual systems which rely on users' interactions, if without a formal data analysis model and explicit objective functions, usually cannot improve the search process in text mining. To bridge the gap between users' subjections and the automatic text analysis process, Guo [8] searches for multivariate clusters in high dimensional data by interactive feature selections accompanied with lower-dimensional projections and computational measurements; Cluster Sculptor [16] interactively explores the hierarchical clustering process of high dimensional data by designing visual representations for various types of data characteristics. The clustering processes in both of the work above are optimized by minimizing distance-based objective functions. However, we realized that users' subjection can be an embedded 'objective function' and some search strategies can be integrated with users' interaction in exploring the feature spaces and searching for a global maximum. For instance, Genetic Algorithm [9], which has a wide range of machine learning and optimization applications [7], is suitable here because of its iterative evaluations.

3.KEYWORDS CLUSTERS REPRESENTATIONS

3.1Text Preprocessing

For text that lacks semantic structures, a keyword is the smallest and most available information unit that can be visualized. After the standard procedures of tokenization and stemming, the unstructured text sequences are transformed into a set of keywords. Stemming could also be done by looking up general purpose external resources such as WordNet [15]. Note that stopping words are also filtered out, and aliases are given a consistent name.

Using the set of extracted keywords, the original texts are labeled using only keywords from this vocabulary. And for each keyword in the restricted vocabulary, there is one occurrence vector per document by labeling the number of occurrences along each line of the document. If the number of document is D, each keyword has a set of occurrence vectors $\{O_{i_l}\}_{l=1}^{D}$, where $(o_{i_l1}, o_{i_l2}, \cdots, o_{i_llp})^T$, and lp is the length of document l.

3.2Clustering Operators

A keywords cluster is a set of keywords denoted by $K = \{k_i\}$. We define two types of clustering operators, KOR and KCO throughout the paper. When keywords are merged with KOR, they are presumably categorized into the same group to suggest a potential semantic-level unit or a concept, so their occurrences in the text data are all dealt with as the occurrences of this concept. Suppose a keyword with occurrence vector $\{O_{i_l}\}_{l=1}^{D}$ is merging with a keyword cluster with occurrence vectors $\{O_{j_l}\}_{l=1}^{D}$using KOR, the occurrence vector $\{O_{k_l}\}_{l=1}^{D}$ of the new cluster is calculated by adding the occurrence vectors of all keywords together with regard to the corresponding document:

$$O_{k_l} = KOR(O_{i_l}, O_{j_l}) = O_{i_l} + O_{j_l}, l = 1,2, \dots D \qquad (3.2.1)$$

. The other type of clustering operator KCO usually suggests the occurrences of a specific event or activity so it groups two keywords clusters $\{O_{i_l}\}_{l=1}^{D}$ and $\{O_{j_l}\}_{l=1}^{D}$ together into a new cluster $\{O_{k_l}\}_{l=1}^{D}$ by counting their co-occurrences in the text:

$$O_{k_l} = (\min(o_{i_l 1}, o_{j_l 1}), \min(o_{i_l 2}, o_{j_l 2}), \cdots, \min(o_{i_l lp}, o_{j_l lp}),\ l = 1,2, \cdots, D \qquad (3.2.2)$$

. In the iterative keywords clustering process, both of the operators are used to merge keyword clusters and produce various shapes.

3.3Weighted Keywords Linked Graph

A keywords cluster can also be viewed as a subgraph of a regular graph that has links between any pair of individual keywords, so the keyword clustering on this graph is inherently a process that searches for the next keyword to connect with a subgraph. We quantify the links between the keywords with a regular graph $G_{kl} = \{V, E, W\}$. V is a one to one mapping from keywords to vertices, where $V(k_i) = v_i$ which is simplified as $\{v_i\}$. E is the set of edges where $\forall\, v_i, v_j \in \mathrm{V}, (v_i, v_j) \in \mathrm{E}$. $W: E \to R$ assigns weight to each edge, indicating the cost of connecting this two keywords. A keyword cluster with keywords $\{k_i\}$ also forms a regular subgraph $H_i \in G_{kl}, H_i = E(v_p, v_q), \forall v_p, v_q \in V(\{k_i\})$.

4.VISUAL PATTERNS EVALUATIONS

4.1Visualization of Keyword Cluster Patterns

The occurrences vectors of keywords clusters provide two specific perceptions for the observers. First, the number of occurrences of a word often indicates the importance or weight of a given topic. Second, the distribution keyword occurrences have over the course of a dialogue often implies a pattern of activity this keywords cluster represents over time. So before the development of visual representation, a threshold OC for keywords occurrences overall data set is used to select keywords with occurrences $\sum_{l=1}^{D} \sum_{j=1}^{lp} o_{i_l j} > OC$ as only keywords with enough significance are considered. Then for each remained keyword in each document, the general shape of keyword occurrences O_{i_l}over time is smoothed into a curve by a Gaussian Filter.

However, one keyword's occurrence vectors of any two documents $O_{i_{l_0}}$and $O_{i_{l_1}}$ do not have one-to-one correspondence as documents run different lengths, therefore normalizing occurrence vectors of all documents using the lengths would result in stretched curve shapes that can not reflect the true accumulated distributions for this keyword. To alleviate this problem, we use a fixed-length Discrete Fourier Transform (DFT) to convert the each time-domain occurrence vector $O_{i_l}, l = 1,2, \cdots, D$ into a fixed length frequency domain intensity vector Of_{i_l} . Each Of_{i_l} is weighted inversely with its total occurrences$\sum_{l=1}^{D} \sum_{j=1}^{lp} o_{i_l j}$, and a weighted average intensities of $\{Of_{i_l}\}_{\{l=1\}}^{D}$ are transformed back to time-domain as the accumulated time-domain pattern for this keyword. In figure 2 one keyword's distributions of individual documents are colored in pink and its accumulated time-domain distribution is colored in purple. The area enclosed by the curve and axes could be filled with color using any polygon-filling method to highlight the patterns. The differing lengths of X-axes in pink pattern windows are the exact lengths of the documents, whereas the length of X-axes in the purple pattern window is the length of FFT transformation. The Y-axes throughout all figures are intensities that are normalized to the range between 0 and 1.

4.2Automatic Evaluation on Shapes of Keywords Cluster Patterns

An iteration of keywords clustering can sample and produce a great number of different keywords clusters patterns. In our automatic evaluation model, a distance function as an 'objective function' provides with a metric to measure the 'goodness' or importance of the geometry of the patterns. Figure 3 shows the main Graphics User Interface (GUI) for IVC on nursing data. The distance function for automatic evaluation is discussed in this section and the interactive evaluation is introduced in later sections.

Users can customize an initial set of templates with interesting shapes to start measuring the shape variations in the cluster pattern population. Those templates are approximated by standard probability distribution functions (PDF), such as $poisson$ distributions, and the pool of templates are dynamic as templates can be added to or removed from the templates during iterations. The bottom row of figure 3 illustrates a few templates in the initial step of evaluation. The X-axes of curves are the length of fixed-length DFT and Y-axes again are normalized intensities ranging from 0 to 1. We define a global function to calculate the distance between cluster patterns and the set of templates by comparing their pattern curves. Denote the patterns space as P.At iteration j, templates are $\{T_i^j\}_{\{l=1\}}^{M}, T_i^j \in P$ and keywords clusters patterns are $\{kp_i^j\}_{\{l=1\}}^{N}, kp_i^j \in P$. The global distance function $D: P \to R$ is defined as:

$$D(kp_i^j) = \min(|kp_i^j - T_{l_0}^j|), \quad l_0 = \max_l \left(RM(kp_i^j, T_l^j)\right), l = 1,2, \cdots, M \qquad (4.2.1)$$

The function RM provides a rough-scale match between the geometries of two pattern curves. Algorithms of patterns matching involving scaling, translation and optimization [19] are available, but they are not used here because they lack the real-time interactivity in our applications. So we propose RM as an approximation to find the best matching between two patterns, and leave the finer-scale matching to users' interaction, because users' perception of shape similarities is usually invariant to scaling and translations. Let kp_i^j be values sampled from function $f_i^j : t \to R$, where $kp_i^j = \{f_i^j(t_0), f_i^j(t_1), \cdots f_i^j(t_p)$ Let $DT(kp_i^j) = \{t_{i_0}, t_{i_1}, \cdots t_{i_p}\}$,where $\frac{df\,{}^\wedge j\_i}{dt} = 0, j = 1,2, \cdots p.$ And let $DTS(kp_i^j) = \{s_{i_0}, s_{i_1}, \cdots, s_{i_p}\}$ where $s_{i_j} = sign\left(\frac{d^{2f}}{dt^2}|_{\{t=t_{i_j}\}}\right), s_{i_j} \in \{1, -1\}, t_i^j \in DT$. DTS is in fact a vector of signs of second derivatives at the positions where the first derivatives are zeros.

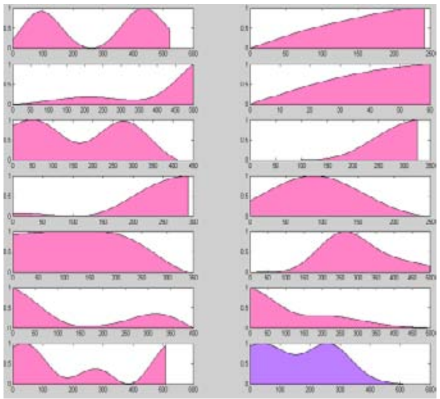


Figure 2. Distributions in individual documents and the accumulated distribution

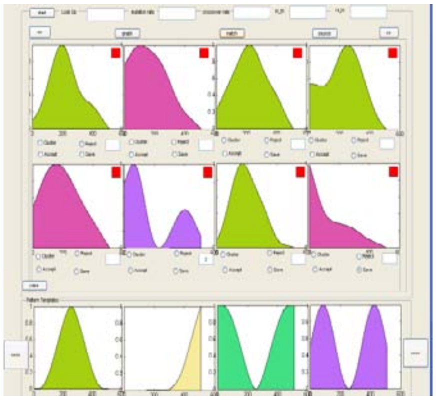
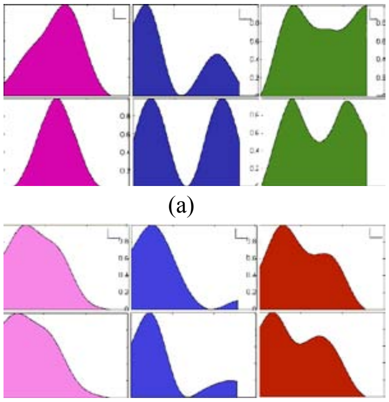


Figure 3. The GUI of iterative visual clustering (IVC)



(a)

(b)

Figure 4. Keyword cluster patterns and their matched templates

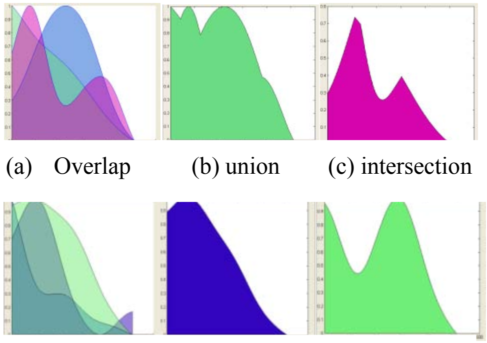
Given $DTS(kp_i^j)$ and $DTS(T_l^j)$, the longest common sequence (LCS) could be found using Dynamic Programming [4]. However, in IVC only the general shapes of the patterns are preserved as occurrences vectors are smoothed, so DTS are often relatively short thus LCS could be found in constant time. Among all the T_l^j found are the longest among all LCS, and the best matched template $T_{l_0}^j$ to kp_i^j is the one that has the minimum Euclidean distance to kp_i^j. Figure 4 (a) (b) display six sets of examples among the best match between cluster patterns in the top rows and the templates in the bottom rows. The same filled color indicates the match. The global distance function also sorts the current generation of patterns in an ascended order to reflect the dispersion among all patterns. A smaller distance value indicates the shape of pattern is close to one of the templates, and a large distance value indicates the shape of pattern is different than all templates. Then the thumbnails of sorted patterns are visualized by two fixed size sliding windows moving from two ends of the sequence towards the middle at the same time. First two rows of thumbnails in Figure 3 are two fixed-size sliding windows. In this way, users have a better overview of the dispersion among the pattern shapes of current population.

5.INTERACTIVE CLUSTERS EVALUATIONS

5.1Interactive Evaluation on Cluster Patterns

We now consider interactive evaluation. Users' opinions on which interesting patterns and which meaningful keyword clusters can also be 'objective functions'. One type of interactive evaluation on cluster patterns is re-matching the patterns and templates to improve the rough-scale match by RM function, as human's perception is known for good at telling shape similarities. This type of interaction re-positions the re-matched pattern in the sorted pattern list based on its new distance, and the improved match accuracy provides a better arrangement on patterns dispersion which will produce better feedback to the next iterations. Another type of interactive evaluation change the templates set. As more dispersed pattern shapes are formed in the search process, users can save a present interesting pattern from current population into templates or remove one useless template when they are navigating the population of patterns. In figure 4(b), three keyword patterns are added into templates so the matched templates for some keyword cluster patterns changed accordingly. The changes in the templates will trigger the global distance function to re-evaluate and re-rank every pattern. The third type of interactive evaluation provides graphical operators on the shape of patterns to help the users compare the geometry of shapes and customize new patterns.

New patterns can be generated by merging more than one patterns using either 'union' or 'intersection' operators. Note that these operators only perform in a graphical level whereby 'union' preserves areas covered by all patterns and 'intersection' only preserves areas that are overlapped among all patterns. Figure 5 shows examples of (a) overlapping of two patterns (b) merging the previous three patterns by 'union' (c) merging the previous three patterns by 'intersection'.



(a) Overlap (b) union (c) intersection

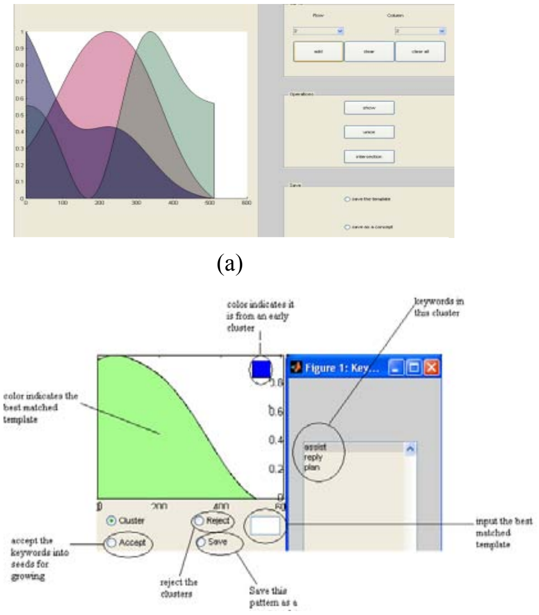
(d) Overlap (e) KOR (f) KCO

Figure 5. Interactive Operations on Patterns and Clusters

5.2Interactive Evaluation on Keywords Clusters

In interactive evaluations, users' can identify keyword clusters that are either meaningless or promising to grow as well. Important keyword clusters are saved as seeds to feedback to the next iteration and meaningless clusters are rejected and removed from current population. Interesting keyword clusters can also be investigated or merged with other clusters by interactively using clustering operators KCO and KOR. Figure 5 (c) –(e) demonstrate the effect of (d) merging two clusters using KCO and (e) merging two clusters using KOR.

Figure 6(a) shows the GUI of graphical and cluster operators on patterns and (b) highlights the different regions that interact with keyword clusters.



(a)

(b)

Figure 6. Visual and interactive evaluation interface

6.KEYWORDS CLUSTER GENERATION

After evaluations, some keyword clusters are deliberately saved or rejected, and the resulted changes in ranking causes the probability of keyword links graph of the current patterns population to be updated. Then how a particular cluster can be grown and which keyword clusters should be grown will be dealt with in the following sections.

6.1Keyword Links Graph Update

Intuitively, the update on weighted keywords links graph is to 'squeeze' the important keyword clusters and 'stretch' the unimportant ones based on their ranking after evaluations. According to the definition of weighted keywords link graph G_{kl} in section 4.2, for every cluster $K_i = \{k_{i_j}\}$ in current population, the weight of the subgraph of K_i is updated by updating every edge in the subgraph as the following equation:

$$E(v_p, v_q) = E(v_p, v_q) * c, \forall v_p, v_q \in V\left(\{k_{i_j}\}\right) \qquad (6.1.1)$$

where c is proportional to the ranking of this keyword cluster among others. Since the importance of keyword cluster patterns decreases from two ends towards the middle, c can be calculated as:

$$c = \left\{\begin{matrix} \frac{i}{CN}, & i \in \left[1, \left\lfloor\frac{CN}{2}\right\rfloor\right] \\ \frac{\left\lfloor\frac{CN}{2}\right\rfloor - i + 1}{CN}, & i \in \left[\left\lfloor\frac{CN}{2}\right\rfloor + 1, CN\right] \end{matrix}\right\} \qquad (6.1.2)$$

where CN is the population size. After updating the keywords links graph, any keywords in important clusters now have smaller weights on their links, so that they are more likely to appear in the further clusters, whereas keywords in unimportant clusters have larger values on their links so that they are unlikely to appear again.

For keyword clusters that are deliberately saved or rejected by users' interactive evaluation, their corresponding subgraphs are updated using equation (6.1.1) by setting c as a constant value less than 1; the subgraphs of rejected keyword clusters are updated using equation (6.1.1) by setting c as a constant value larger than 1.

6.2Population Control

After updates take effect, the unimportant patterns in current sorted population are filtered out by setting up two quantile threshold indices lo and hi in the rank list of patterns. Three candidate sets, including **seeds for growing, seeds for mating and seeds for mutation,** are also generated from sorted population to produce new populations to reproduce and fill the gap caused by filtered out clusters .

The number of clusters in each seeds set is determined in a way to fix the size of population in each generation. The following equation is used to circumscribe the parameters in population control:

$$rn + hi - lo + 1 = sg + \frac{sm}{2} + \left(sg + \frac{sm}{2}\right) * murate \qquad (6.2.1)$$

where rn is the number of rejected clusters, lo is the low index end of the quantile filter, hi is the high filter threshold, sg is the number of clusters kept for growing or seeds for growing, sm is the number of seeds for mating, and $murate$ is the mutation rate. Note that the number of saved clusters sg and the number of rejected clusters rn are both interactively adjustable and also in-directly changes other parameters in the equation.

6.3Seeds for Growing

Seeds for growing, kept by users, expand themselves in the next iteration. We assume that the probability of one keyword being selected is inversely proportionally to the weight of the link between the keyword and the cluster. Intuitively, a 'nearer' neighboring keyword of the cluster is more likely selected as compared to the 'further' ones. If the link between a specific keyword node $v_i$to a keywords clusters $V\left(\{k_{i_j}\}\right)$ connects the keyword and the center of H_i, the weight of the link between one keyword node v_i and the keywords clusters $V\left(\{k_{i_j}\}\right)$ is $\frac{1}{n}\sum_{v_p \in v_{i_j}} W(v_p, v_i)$. The links from keyword clusters $\{v_l\}_{l=1}^{n}$ to each keyword node $v_i$are sorted in an ascend order , and the probability of selecting a particular link is assigned as $p_i = \frac{LN-i+1}{\sum_{i=1}^{LN} i}$ using Roulette Wheel rank weighting [19], where i is the ranking of the link among the others, LN is the number of links. A random number r is drawn from uniform distribution between zero and one, the link i in the sorted link list with $cp_{i-1} < r < cp_i$is selected, where $cp_i = \sum_{j=1}^{i} p$. It is easy to prove that $p(r \le cp_i, r > cp_{i-1}) \propto p_i$ when r has a uniform distribution.

Once a seed is grown, the occurrence vectors of keyword clusters are also updated according to equation (4.2.1). The weight of subgraph H_i of the keywords clusters also is increased by adding more edges into the subgraph. Every cluster in the **seeds for growing** keeps growing until the total weight of the subgraph exceeds the weight threshold of current iteration. So a better cluster whose subgraph was 'squeezed' previously will grow faster and an unsatisfying cluster whose subgraph was 'stretched' will grow slower.

6.4Seeds for Mating

The keyword clusters in seeds for mating are also selected by Roulette Wheel weight ranking, where the probability of each keyword cluster being selected is proportional to their rank in current population sorted by visual evaluation. Mating are expected to introduce newer shapes, so patterns at both end of current sorted population have higher probability of being selected into the mating group. Thus the sorted population is divided into two ranges
$\left[1,\left\lfloor\frac{CN}{2}\right\rfloor\right]$and $\left[\left\lfloor\frac{CN}{2}\right\rfloor+1,CN\right]$, where CN is the number of current population. Each keyword cluster in two divisions is assigned a probability $p(K_i)$ of being selected by the following equation:

$$p(K_i)=\begin{Bmatrix}\frac{\left\lfloor\frac{CN}{2}\right\rfloor-i}{\sum_{i=1}^{\left\lfloor\frac{CN}{2}\right\rfloor}i}, & i\in\left[1,\left\lfloor\frac{CN}{2}\right\rfloor\right]\\ \frac{i}{\sum_{i=1}^{CN-\left\lfloor\frac{CN}{2}\right\rfloor}i}, & i\in\left[\left\lfloor\frac{CN}{2}\right\rfloor+1,CN\right]\end{Bmatrix} \qquad (6.4.1)$$

Then seeds for mating are evenly sampled in two separated passes running on the two intervals, and the keyword clusters from two runs are paired and merged using KOR clustering operator in the order from top to the bottom. The weight of the newly generated keywords clusters and the occurrence vectors of keyword clusters are updated accordingly.

6.5Seeds for Mutation

Seeds for mutation are randomly sampled, unique keyword clusters from the new clusters generated by seeds for growing and seeds for mating with mutation rate $murate$. The mutated keyword cluster is resulted from flipping any KOR clustering operators into KCO, which represents the co-occurrences among all keywords in this cluster. The weight of keywords clusters remains the same while the occurrence vectors $\left\{O_{i_l}\right\}_{l=1}^{D}$ of the mutated cluster is updated according to equation (3.2.1) and (3.2.2).

To this end a new generation of keywords clusters is formed, they are visualized and evaluated again.

7.RESULTS AND DISCUSSIONS

7.1The Nursing Data Application

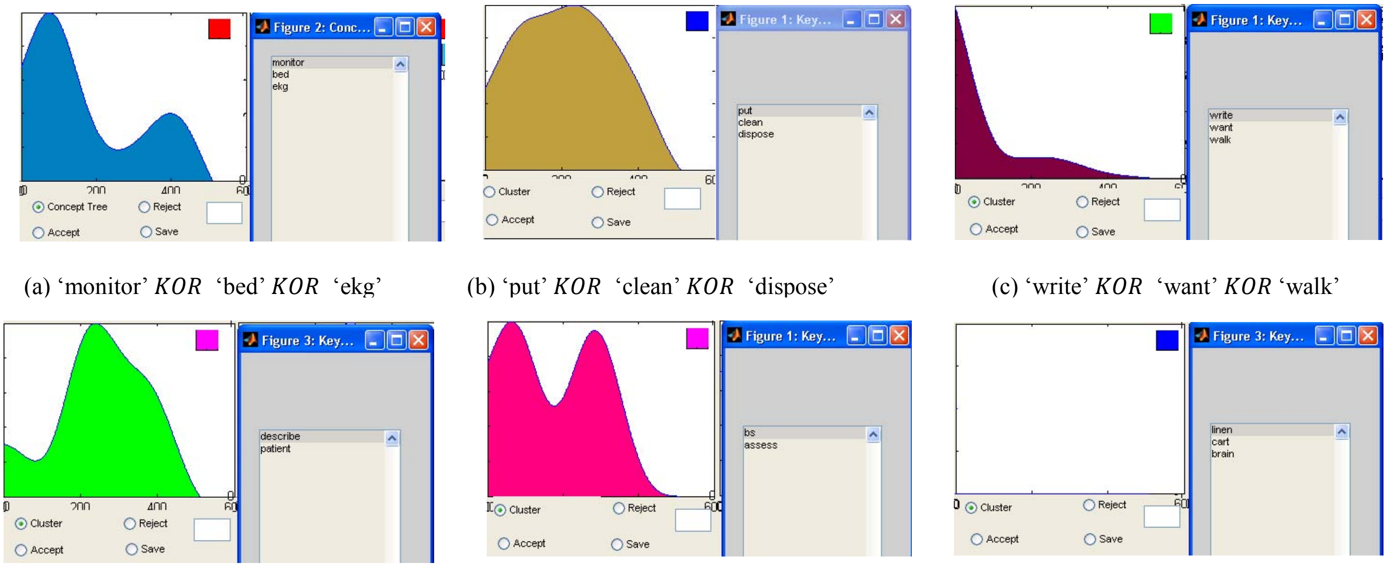
A procedure for manual recording of direct observations of Registered Nurses (RN) work was developed to help working nurses and nursing domain experts to investigate and understand the numerous activities RNs engaged in to manage the environment and patients flow. Observation data was recorded on legal pads, line by line, using an abbreviated shorthand method as unstructured text. A segment of a sample session of nursing narratives is shown below:

- Walks to Pt #1 Room
- Gives meds to Pt #1
- Reviews what Pt #1 receiving and why
- Assesses how Pt#1 ate breakfast
- Teaches Pt#1 to pump calves while in bed
- Explains to Pt#1 fdngs- resp wheezes
- Reinforces use of IS Pt#1
- Positions pillow for use in coughing Pt#1
- …………………

What to the outside casual observer appears to be single task elements becomes a much more complicated array of overlapping functions with inter-related patterns and trends. There are two basic anticipated applications to analyzing RN work: (1) identification of work patterns related to non-clinical work or basic nursing work (2) staffing and assignment implications based on work patterns across time. We applied IVC to 13 data sessions of nursing narrative texts. After preprocessing the text with procedures described before, 104 keywords, about 20 percent of all keywords, are resulted and each keyword is represented by 13 vectors corresponding to each session. Different sessions have differing running lengths and are neither synchronized nor aligned. Therefore we used the accumulated time-domain distributions previously introduced to rule out the synchronization problem.

7.2Results of Keyword Clusters

During the iterations, various shapes other than templates are introduced and two general types of keyword clusters can be derived to demonstrate higher level knowledge, regarding to the nursing working pattern and environment. The first type is the **general behavior patterns** of nurses that can be implied keyword clustered using KOR. Figure 7 (a)-(c) are three examples of this cluster type, and a small colored square on the right top classifies the sources of a pattern: surviving the evaluation (red), from one seed growing (blue) and from seeds mating (green). The keywords cluster on the right of (a) is represented by the curve on the left. 'monitor', 'bed' and 'ekg' are all equipments in patients' room, implying nurses are staying in ekg rooms and operating on the equipments in a timely way the pattern of the curve indicates. With one big peak and a much smaller peak sitting at two ends of the time line, nurses may have 'rushing hours' in operating equipments in patients' room, both at the beginning and the end of the day. As the first peak is much larger than the second, it may indicate the need for operations decreased severely or operating on equipments in patient's room are usually arranged in early hours of the working session. Given the span of two peak areas are nearly the same, those jobs may be routine so that nurses can finish them in a fixed period. Figure 7(b) indicates the working pattern of nurses' clearing up equipments and (c) shows the working pattern of a group of actions including writing documents, walking among places or listening to patients' wants The second type of concept is an **event pattern** or **assignment pattern** represented by co-occurrences among keywords. A small cyan square on the right up indicates that this group of keywords is from mutation and has clustering operators KCO merge them together. In figure 7 (d), the appearances of peaks on the left in fact represent the co-occurrences of the keywords meaning that nurses listening to the patients describe their syndromes and so on. A large hump sits in the middle of the time line indicating it is mostly in the middle of the working session that the nurses get patients' own descriptions about their situations. This working pattern may either response to the shuffle-in of patients during the middle of the working session, or show nurses are scheduled to answer questions to patients in bed; the pattern in figure 7(e) indicates the nurses' activity of checking brain sheets for orders; the pattern in figure7 (f) has zero intensity everywhere thus demonstrates that nurses do not access linen room, carts and brain sheets at the same time.



(a) ‘monitor’ KOR ‘bed’ KOR ‘ekg’

(b) ‘put’ KOR ‘clean’ KOR ‘dispose’

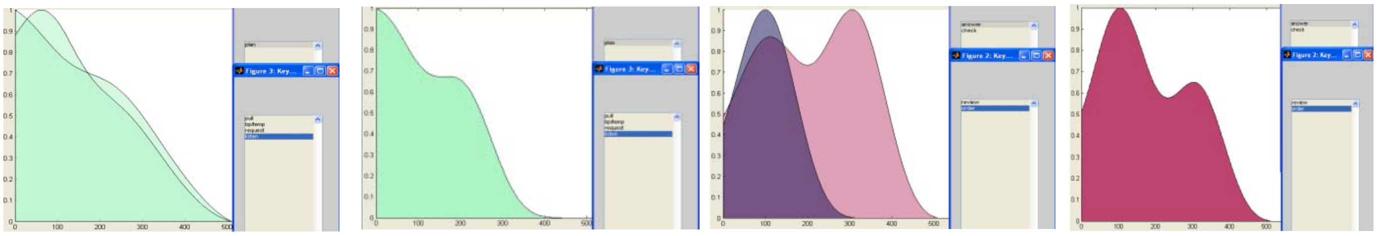
(c) ‘write’ KOR ‘want’ KOR ‘walk’

(d) ‘describe’ KCO ‘patient’

(e) ‘bs’ KCO ‘assess’

(f) ‘linen’ KCO ‘cart’ KCO ‘brain’

Figure 7. Keyword clusters that imply working patterns of nurses



(a) ‘answer’ KOR ‘check’ overlap with ‘review’ KOR ‘order’

(b) ‘answer’ KOR ‘check’ KOR ‘review’ KOR ‘order’

(c) ‘plan’ overlap with (‘pull’ KOR ‘bp/tmp’ KOR ‘request’)

(d) ‘plan’ KCO (‘pull’ KOR ‘bp/tmp’ KOR ‘request’)

Figure 8. Keyword clusters that represent complex working patterns

Overlaying patterns together and coloring them with transparency can also obtain insights of nursing tasks as shown in figure 8(a). A specific assignment of ‘answer’ KCO ‘check’, which represents answering questions and regular checking for patients, has the first peak in pattern almost coincided the first peak in the pattern of the assignment of ‘review’ KCO ‘order’ (reviewing orders). If the two assignments usually happen in different places around nursing working environment, then there might be a schedule conflict or at least an inefficient arrangement in nursing tasks. To further investigate interleaving behavior patterns or nursing events, more complicated keyword clusters can be interactively constructed using cluster operators KCO and KOR. Figure 8(b) shows the example of merging results of two patterns in figure 8 (a) as (‘answer’ KCO ‘check’) KOR ‘review’ KCO ‘order’. Therefore the unified pattern of two event patterns, if considered as a general behavior pattern, is generated using interactive cluster operators. Similarly, pattern (‘pull’ KOR ‘bp/temp’ KOR ‘request’ KOR ‘listen’) and pattern ‘plan’ overlapping in figure 8(c) have the pattern of figure 8 (d) if they are considered as two necessary parts of one event pattern and merged by cluster operator KCO . Sophisticated concepts therefore have a explorations, and would be helpful in improving the design of the working environment

8.CONCLUSIONS AND FUTURE WORK

The iterative visual clustering (IVC) we proposed in this paper aims to find semantically meaningful keyword clusters and their patterns over time line from unstructured text data by searching the possible combinations of keyword clusters, and supporting visual and interactive evaluations and interpretations. Compared to conventional text mining methods, our IVC does not define any explicit objective functions but is driven by users’ intentions in a stochastic search process to narrows down the search space of the huge number of keywords cluster combinations. Users’ intentions are feedback by interactions which both inject users’ domain knowledge and direct the search process.

We will further refine the stochastic model of IVC and enhance the visualizations and interactivity of IVC. To thoroughly explore the inherent knowledge rooted in nursing narratives, we are now expecting larger data sets. Meanwhile, IVC is applied to other free running text corpus, such as biomedical literatures, to test their effectiveness as a visual data mining framework

9.REFERENCES

[1] Akaishi, M., Hori, K. and Satoh, K. Topic Tracer: a Visualization Tool for Quick Reference of Stories Embedded in Document Set. In *Proceedings of the Tenth International Conference on Information Visualization, 2006.* (*IV'06)* (London, UK, July 5-7, 2006). IEEE Computer Society, Washington, DC, 2006, 101-106.

[2] Cafias, A.J., Carff, R., Hill, G., Carvalho, M., Arguedas, M., ESKRIDGE, T.C., Lott, J. And Carvajal, R. Concept Maps: Integrating Knowledge and Information Visualization. *Knowledge And Information Visualization: Searching for Synergies*, Springer Berlin / Heidelberg, 3426(2005), 205-219.

[3] Chen, C. And Kuljis, J. The rising landscape: A visual exploration of superstring revolutions in physics. *JASTIS 54*, 5(2003), 435-446.

[4] Cormen, T.T., Leiserson, C.E. and Rivest, R.L. *Introduction to algorithms*. MIT Press Cambridge, MA.1990.

[5] Delest, M., Munzner, T., Auber, D. and Domenger, J.P. 2004. Exploring InfoVis Publication History with Tulip. *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'04)* (Austin, TX, Oct10-12,2004), IEEE Computer Society, 2004.

[6] Feldman, R., Dagan, I. and Hirsh, H. Mining Text Using Keyword Distributions. *Journal of Intelligent Information Systems 10*, 1998, 281-300.

[7] Goldberg, D.E. And Holland, J.H. Genetic Algorithms and Machine Learning. *Machine Learning 3*,1988 , 95-99.

[8] Guo, D. Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization 2*, 2003, 232-246.

[9] Haupt, R.L. and Haupt, S.E. *Practical Genetic Algorithms*. Wiley-Interscience, 2004.

[10] Havre, S., Hetzler, E., Whitney, P., Nowell, L., Div, B.P.N. and Richland, W.A. ThemeRiver: visualizing thematic changes in large document collections*. IEEE Transactions on Visualization and Computer Graphics,* 8(2002), 9-20.

[11] Liu, H. and Maes, P. Rendering Aesthetic Imparessions Of Text In Color Space. *Interational Journal on Artifical Intelligent Tools,* 15(2006), 515.

[12] Mamykina, L., Goose, S., Hedqvist, D. and Beard, D.V. 2004. CareView: analyzing nursing narratives for temporal trends.*In CHI '04 extended abstracts on Human factors in computing systems*, (Vienna, Austria, April 24 - 29, 2004), ACM Press, New York, NY, 2004, 1147-1150.

[13] Manning, C.D. and Sch`eutze, H. *Foundations of Statistical Natural Language Processing.* MIT Press, 1999.

[14] Marks, J., Ruml, W., Ryall, K., Seims, J., Shieber, S., Andalman, B., Beardsley, P.A., Freeman, W., Gibson, S. and Hodgins, J. 1997. Design galleries: a general approach to setting parameters for computer graphics and animation. In *Proceedings of the 24th Annual Conference on Computer Graphics and interactive Techniques* .(Los Angeles, California, August 3-8,1997), ACM Press/Addison-Wesley Publishing Co., New York, NY, 1997,389-400.

[15] Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K.J. Introduction to WordNet: An On-line Lexical Database*. *International Journal of Lexicography*,3(2004) 235-244.

[16] Nam, E.J., Han, Y., Mueller, K., Zelenyuk, A. and Imre, D. ClusterSculptor: A Visual Analytics Tool for High-Dimensional Data. Visual Analytics Science and Technology. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology, 2007. (VAST 2007)*(Sacramento, California, Oct28-Nov1, 2007). IEEE Computer Society, Washington, DC, 2007, 75-82.

[17] Plaisant, C., Mushlin, R., Snyder, A., Li, J., Heller, D. and Shneiderman, B. LifeLines: Using Visualization to Enhance Navigation and Analysis of Patient Records. *The Craft of Information Visualization: Readings and Reflections*. Morgan Kaufmann, San Francisco.2003.

[18] Rohrer, R.M., Ebert, D.S. and Sibert, J.L. 1998. The Shape of Shakespeare: Visualizing Text using Implicit Surfaces. In *Proceedings of the IEEE Symposium on Information Visualization 1998.(InfoVis'98)*(Research Triangle Park, NC, Oct 19-20,1998), IEEE Computer Society, Washington, DC,121-129.

[19] Veltkamp, R.C. and Hagdoorn, M. *State-of-the-art in Shape Matching*. Univ.; Niedersächsische Staats-und Universitätsbibliothek.1999.

[20] You, Q., Fang, S. and Ebright, P. Visualizing Unstructured Text Sequences Using Iterative Visual Clustering. *Lecture Notes in Computer Science* 4781(2007), 275-284.

[21] Zhu, W. and Chen, C. Storylines: Visual exploration and analysis in latent semantic spaces. *Computers & Graphics,* 31(2007), 338-34.