Digital Library Information Categorization, Visualization, and Retrieval

Jim X. Chen, Ken Alford and David Grossman
Department of Computer Science
George Mason University

Ophir Frieder
Department of Computer Science
Illinois Institute of Technology

Abstract

We present a new information categorization, visualization, and retrieval method. First, we map each document in a digital library into a vector which contains the frequencies of all words in the entire collections stripping off those strings which do not contribute distinctive information, then we correlate each vector with a hyper-pixel in a 2D space. We call a pixel a hyper-pixel because it can be used as a handle to retrieve the corresponding information. The hyper-pixels are categorized by its corresponding vectors using a physically-based method. Pixels adjust their current locations according to an interactive force field. A library of documents will evolve into a graphics hyper-image which can be visualized in multiple layers and levels of detail. The hyper-image displays the locations and distances among the documents. It provides the local density and clustering of the documents. Information and documents can be retrieved from this hyper-image through the graphical user interface according to the spacial location, area, or keywords specified by the application. No priori information about document content is required. This method provides a different means for information categorization and retrieval. It provides the application user a visualization of the digital library system in different levels of detail and categorization schemes, and retrieves the information according to user preference after visualization.

Keywords: information, categorization, retrieval, visualization, physically-based, graphical user interface, hyper-image, document, library

1 Introduction

Sorting and categorizing the enormous volume of text now available in machine-readable digital libraries has become a pressing problem. Information description, visualization, and retrieval is the reverse process and is equally challenging. Recently, digital library information visualization and retrieval using graphics and graphical user interface were studied in different perspectives. Korfhage [11] used vector distances to represent the similarity of the documents. The problem with his method is that the information is categorized by reference points. When the reference points change, which occurs with the change of queries, the distances between the reference points and all the documents in the system have to be recalculated, which is very computationally expensive. Chalmers and Chitson [1] used particles in 3D space and potential fields, and defined the distance between particles to model the relationships between the various bibliographic data. Chimera [2] discussed using value bar to represent combinations of information visualization and navigation for multi-attribute listings. Eick, Sumner, and Wills [6] discussed visualization of the results of keyword and author database searches. The idea is to display the hits on a journal-by-year grid, using size and color to show the number and type of hits for each symbol. Hemmje [9] introduced a 3D user interface for information retrieval. It
provides a human computer interface for navigation and demonstration of the information system. Robertson, Card, and Mackinlay [13] introduced an information visualizer which explores a user interface paradigm. Spoerri [15] introduced a representation that can be used as a visualization tool as well as a visual query language to help users search for information. Kjell, Woods, and Frieder [8] studied how to discriminate authorship using visualization. They categorized documents using the frequencies of the two-tuples. The representation for each prototype text, or any of the individual document, is a vector of 26² features. Then, they used the Karhunen-Loeve method [7] to transform a feature vector into 2D coordinates, which determines a point in an image. Their purpose was to compare the images to determine the authorship of the documents. The problem is that the Karhunen-Loeve method is very time consuming. Damashek [4] used n-grams to categorize language-independent text. For a 5-gram English text, the possible number of items to be counted for is at least 26⁵, which is impractical for a complete array representation. However, most of the 5-grams are never encountered. Huge reserves of computer memory for n-gram statistics are therefore unnecessary. Instead, document vectors can be stored conveniently by indexing each n-gram in a consistent manner using hash algorithm [10]. Possible collisions were ignored. Chen and Ophir [3] borrowed ideas from [4, 8] and introduced the idea of hyper-image. However, their information categorization is very slow due to the Karhunen-Loeve method.

We present a new information categorization and retrieval method using a physically-based approach for information categorization and a graphical-based approach for visualization and retrieval. The method presents the information system as an image, providing the users with a picture of the clustering and categorization of the system documents, and allowing the users to retrieve documents by specifying a region of information. In practical application, this is a totally different method of library system interface, better for users who want to access and retrieve a range of related information without exactly matching the complicated combinations of keys. We differ from all at the above by that we have a clear visualization of the digital library system as well as levels of detail with flexible information specification and retrieval. There is no additional calculation or traversal of the information clustering once it is set up.

Our new method borrows ideas from [1, 3, 4, 8, 11]. First, we map each document into a vector which contains the frequencies of each words after stripping off those strings which do not provide distinctive information. We use words instead of n-grams because any n-gram system contains meaningless strings, which do not contribute to the information categorization, and in return, to the visualization and retrieval. Although the number of possible words are so large that a complete list is beyond practicality, only a small number of words will be most probably encountered. More importantly, only a few of them will appear frequently, which is indexed using a hashing algorithm.

After transforming a document into a vector, we correlate it with a hyper-pixel in a 2D space within a graphical user interface. We call a pixel a hyper-pixel because it can be picked up and used as a handle to the corresponding vector, and in turn, to the corresponding document. The hyper-pixel will move to its destination in the 2D space through a physically-based modeling method [16], which will be discussed in Section 2 The cosine similarity measure [14] is commonly used in information retrieval to compare the differences between vectors. The physically-based modeling approach will dynamically move the similar documents closer and the dissimilar documents farther in distance. A library of documents forms a graphics hyper-image which can be visualized in multiple layers. The locations of the hyper-pixels in the hyper-
image correspond to the vectors after the physically-based modeling process. A graphics hyper-image is, therefore, on one hand, a visualization of the whole information system, and on the other, a reference to retrieve relevant documents. Without loss of generality, here we interchangeably use hyper-pixel and pixel, hyper-image and image.

Information and documents are retrieved from this hyper-image according to the spatial location, area, or keywords specified by the user through the graphical user interface. The similarity between two hyper-pixels are inversely proportional to the distance between them. Therefore, if we specify an information vector, it corresponds to a hyper-pixel in the hyper-image. If we choose the vicinity area of that pixel, we can retrieve related documents. Given a few key words and their weights (optional), we can locate their corresponding vicinity, and retrieve certain number of related documents according to the shape and size of the area surrounding the key point. In addition, we can choose certain points and translate them to a same location, and retrieve the information in combination of these points.

No priori information about document content is required. This method provides a different means of information categorization and retrieval. One contribution of this system is that it provides the user with a better picture how the system documents are distributed and correlated. The user, based on an image, chooses and retrieves the related information. It provides the application user with a visualization of the information system in different levels of detail, and it retrieves the information according to user preference after visualization. Once the pixels are chosen, the retrieval process is direct without time-consuming search and find.

2 Categorization

An entire document can be represented as a vector whose components are the relative frequencies of its distinct words. Each document is processed by converting each word into an indexing key, ordering a vector by key values, counting and storing the number of occurrences of the same key value, and dividing the number of occurrences of each distinct key by the final total number of keys. Upper case is converted into lower case, and non-informative words and symbols (such as punctuations, spaces, ‘a’, ‘be’, ‘the’, etc.) are ignored by checking its corresponding index.

Different documents have different frequency distributions. Documents dealing with similar topics tend to have high frequencies in the same keys. Figure 1 shows a section of the vector frequency distributions of two papers of related topics. The basic assumption is that two documents whose distributions are “similar” in some useful sense are likely to deal with related subjects.

![Fig. 1: vector similarity of documents dealing with the same subject](image)
Let a document contain $m$ distinct words with $u_l$ occurrences of word index $l$. Then the weight assigned to the $i$th vector components will be

$$freq_l = \frac{u_l}{\sum u_j}$$

(1)

where

$$\sum_{j=1}^{m} freq_j = 1$$

(2)

The cosine of the angle between two document vectors $i$ and $j$ is

$$cos \theta = \frac{\sum_{l=1}^{m} freq_{i,l} freq_{j,l}}{\sqrt{\sum_{l=1}^{m} (freq_{i,l})^2 \sum_{l=1}^{m} (freq_{j,l})^2}}$$

(3)

equation (3) is the cosine similarity measure commonly used in information retrieval. Co-linear vectors will have a cosine of 1.0; dissimilar vectors will have a very small cosine. The score given by equation (3) is also the cosine of the angle between two vectors in the high-dimensional document space, as viewed from the absolute origin [4].

We use a physically-based modeling approach for the movements and correlations of the pixels in the 2D space. We build up a force field in the image and each pixel moves to its corresponding location by the force generated according to its relation with other documents. Let’s assume two document vectors are represented in linked lists as follows in Figure 2. The first indexing key in document 1 (doc1) equals the first indexing key in document 2 (doc2), which is a factor contributing to the similarity of these two documents. The second indexing key in doc1 has no corresponding indexing key in doc2, which is a factor contributing to the dissimilarity of the two documents. Therefore, we can generate a combination of forces between the documents. The factors contributing to the similarities will generate positive forces which attract them together, while the factors contributing to the dissimilarities will generate negative forces which repel each other apart. All the document pixels are correlated in the 2D space by their corresponding forces. They move towards the locations where the vicinities are more related documents.
Now let's consider the force in $x$ direction between document $i$ and $j$. For all the keys in doc$_i$, if $key_{i,k}$ has a corresponding key in doc$_2$, then it generates a positive force inversely proportional to the differences between their relative frequencies of the indexing key. The force should also be proportional to their distance between the two document pixels so that they will move faster towards each other if they are far apart. So we have a component as follows:

$$f_{x_i,j} = \sum \frac{d_{x_i,j}}{|freq_{i,k} - freq_{j,l}| + \varepsilon} \quad \text{for all } k, l, \text{ if } key_{i,k} = key_{j,l}$$

(4)

where $\varepsilon$ is a small number preventing the denominator from being zero. For all the keys in doc$_i$, if $key_{i,k}$ has no corresponding key in doc$_2$, then it generates a negative force proportional to the frequency of the indexing key. The force will be significant when the two document pixels move close to each other. So we have a component as follows:

$$f_{x_i,j} = -s\sum freq_{i,k} \quad \text{for all } k, \text{ if } key_{i,k} \text{ has no corresponding key in } doc_j$$

(5)

where $s$ is a scale factor adjustable for different levels of detail. Therefore, assuming there are $n$ documents, the force field generated for document $i$ is as follows:

$$F_{x_i} = \sum_{j=1}^{n-1} (f_{x_i,j} + f_{x_i,j}^\prime)$$

(6)

The corresponding force equations in $y$ direction are as follows:

$$f_{y_i,j} = \sum \frac{d_{y_i,j}}{|freq_{i,k} - freq_{j,l}| + \varepsilon} \quad \text{for all } k, l, \text{ if } key_{i,k} = key_{j,l}$$

(7)

$$f_{y_i,j} = -s\sum freq_{i,k} \quad \text{for all } k, \text{ if } key_{i,k} \text{ has no corresponding key in } doc_j$$

(8)

$$F_{y_i} = \sum_{j=1}^{n-1} (f_{y_i,j} + f_{y_i,j}^\prime)$$

(9)
Examining equation (4) to equation (9), we can see that the only variables are the distances in $d_{xi,j}$ and $d_{yi,j}$. $s$ is an adjustable constant, and all the rest are existing constants.

According to newton’s second law $F = ma$, if we assume each pixel has $m=1$, we have $a = F$. Thus we can use conventional physically-based modeling and simulation method to calculate the movement of each vector pixel in the 2D space [16]. The system will iterate as follows:

```plaintext
while (loc.new(i) ≠ loc.old(i)) do following
  0. Reset the location of current pixel: loc.old(i) = loc.new(i);
  1. Calculate F(i) using equation (4) through equation (9);
  2. Acc(i) = F(i);
  3. Calculate the current velocity: vel(i) = vel(i) + Acc(i)*Dt; \textit{where Dt is a time constant}
  4. Calculate the new location of the pixel: loc.new(i) = loc.old(i) + vel(i)*Dt;
end of while
```

The spacial distances between the pixels vary reflecting their similarities and differences. After a certain period of time, the system will achieve a balanced state when there is no pixel movement any more. Thus the iteration stops. Therefore, documents are categorized as hyper-pixels in a 2D space, and the distances between pixels are the measurement of the similarities between the documents.

3 Visualization

After the physically-based categorization, a set of documents result in a 2D hyper-image. The hyper-image illustrates the distribution and clustering of the documents according to their corresponding features. We can specify any location or area so that the system can provide the predominant features corresponding to that area, and the number of documents covered in that area. For example, the Wall Street Journal portion of the Tipster collection used in the NIST Text Retrieval Evaluation Conference (TREC) can be transformed into an image. Figure 6 represents all 74,520 (1990-1992) Wall Street Journal articles used in the TREC-4 competition. The total Tipster collection consists of about 2 gigabytes of text (over 1 million documents) including articles and documents from the Wall Street Journal, Associated Press, Department of Energy, San Jose Mercury News, Ziff Communications Company, US Patent Office, and Federal Register. Because of the large number of documents, we can scale a certain area and zoom-in to have a better visualization of certain information of interest. Figure 7 is the zoom-in of the area around Merger.

Our Graphics User Interface (GUI) allows many different operations for categorization, visualization, and retrieval. The system has an iconic menu and a menu bar with pulldown options. The iconic menu provides different ways (rectangle, ellipse, freehand) of selecting a range of information, corresponding to the area specified in the image. It provides also Translation (T), Merge (+), Zoom-in, Zoom-out, and view in multiple layers (M) of the specified area. Under the System menu, you can use New to start a new information system on a per user basis (similar to a view in the relational database model), Add to add a new document into the system, Save to save this information correlation structure, and Quit to exit from the system. After looking at the hyper-image, the user can specify an area, or a few keywords using GUI, then the system can provide the number of documents covered in the selected area, most
dominant key words in this area (the number is user specified), and titles associated with the
covered documents, etc., according to the user specification. Under the Queries menu, you can
use Subject, Keywords, Authors, or others to locate the corresponding information in the image.
Under the RetInfo menu, you can use Documents (Number, Titles, Authors, Search) to check the
number, titles, authors or search results of the documents in the specified area, you can use
Keywords (<100, <50, <20, or Search) to check the most frequently appearing keywords
(confined in the given numbers, as shown in Figure 7). Under the Utilities menu, you can Merge
specified areas and view the system in Multiple layers.

By scaling (i.e., zooming-in on) a certain area, we can view different levels of detail concerning
this specific information. It is easier to subdivide and select information. This process can be
repeated so that we can have different layers of information representation and levels of detail.
Figure 3 is an example of representing different levels of detail, and Figure 7 is an example of
our zooming-in around Merger in the Wall Street Journal portion of the Tipster hyper-image.

![Figure 3: different levels of detail in information visualization](image)

By shifting the center of dominant features, we can visualize the system in a different
perspective. For example, if we retrieve features corresponding to two different important key
words, we can shift them into the same origin. This way the center of information includes two
features as shown in Figure 4.

![Figure 4: document subject Stock under Merger and Break-up](image)

4 Retrieval

From the corresponding document vectors we can retrieve and display the most prominent key
words, which represent the main stream information of the chosen documents. On the other hand,
we can specify keyword vectors to find corresponding pixels in the picture, and by clicking on a
point or a set of points, we can obtain related information. When we display a chosen area, we
can display the number of pixels chosen. If the number is too big, we can use the levels of detail to amplify the chosen area, and refine the next selection from the amplified area. We can shift centers of remote areas to merge different topics and retrieve sub-topics from different categories. For example, we can translate documents concerning Merger and Break-up to the same location, then we can choose an area from the merged image, so we can access Stocks in both Merger and Break-up specifically (Figure 4).

Users select featured documents from the system by different means provided in the GUI. Users can specify a set of documents by giving keywords, subjects, or other information to refine the selection. Relevance feedback is an underlying premise of the system and is directly supported. An area around the specified points contains related information. Each pixel in the hyper-image is a handle to its corresponding vector, and it in turn, to its corresponding document. By specifying a region, a set of related documents is chosen. The information retrieval abstract data structure is described in Figure 5. As shown in the figure, each pixel is a handle to its corresponding vector, and therefore to its corresponding document. Thus the information retrieval is a direct process without time-consuming search.

**Conclusion**

We have introduced our research in physically-based interactive information categorization and graphical-based hyper-image information retrieval. The advantages of this method include that the physically-based method guarantees the information categorization reflects the spacial correlations of the documents, the resulting hyper-image provides the user a visualization of the information system, the image allows users to retrieve information in different levels of detail, and the retrieval of the related information takes no extra effort and time because of the structure of the hyper-image system. User select domains of related information instead of matching keywords exactly. We anticipate this system to be used in library system management and to provide readers with a better approach to access and retrieve information.
1 References


Fig. 6: information hyper-image clustering

Fig. 7: zoom-in on Merger and return information about keywords