Semantic Illustration Retrieval for Very Large Data Set

Song Kai[1] , Huang Tie-Jun[2] , Tian Yong-Hong[1]

[1]Digital Media Lab, Institute of Computing Technology, Chinese Academy of Sciences
Beijing, 100080, P. R. China

[2]Institute for Digital Media, Peking University
Beijing, 100871, P. R. China

Abstract **— In this paper, we present a retrieval system that performs the illustration retrieval on very large data set. The traditional text-based retrieval systems often perform poorly on the illustration retrieval, because some illustrations are uncaptioned. Even worse, textual information are often mixed with noisy information and therefore fail to represent the illustrations accurately. To overcome the problem, we propose a semantic model for illustration retrieval. In this model, we first extract the shape information whose similarities are then used to construct two link graphs. Based on the graphs, we execute the auto-captioning procedure on the uncaptioned illustrations. In addition, cross-modal analysis is applied to get rid of the noisy information and reduce the dimensionality of the feature vectors. Finally, we introduce a re-rank scheme that returns as many subtopics related to the query as possible along with the improvement in relevance. Experiments on approximately 500,000 illustrations showed that our system performs efficiently in retrieving the illustrations with high relevance and diversity.**

Index Terms **— illustration retrieval, feature extraction, cross-modal analysis, ranking**

I. Introduction

As the digital library develops rapidly, a large number of illustrations in digital books are available to the users. Journalists and publishers need to query proper illustrations for newspapers, books and articles. Historians and archaeologists may need illustrations in ancient books to verify their findings. Therefore, it is very important to develop techniques for users to find the pictorial information they want from digital libraries.

In [3], Cohen and Guibas proposed an illustration retrieval system based on visual features. Shape information is extracted for indexing and retrieval. Nevertheless, the method is mainly focuses on the computer-generated technical illustrations which are mainly composed of graphics primitives (lines or circular arcs, marks, etc.). Therefore, the method can not be generalized in other complicated illustrations such as images of people, animals and landscapes. Previous work in [3] mainly focuses on visual features of illustrations and therefore fails to explore the semantics of illustrations accurately. Librarians usually rely on text-based retrieval using their bibliographical catalogue as reliable metadata, but manual attachment of textual metadata to the illustrations are too time-consuming. In [6][7], Hu *et al* proposed several methods to extract titles and other textual information from various documents automatically. However, titles or captions of illustrations are often missing and some of the collateral texts are irrelevant to the content of the illustrations and therefore carry lots of noisy information. To overcome the problem, a method to structure the visual content of digital libraries using Content-Based Image Retrieval (CBIR) is proposed in [1]. By grouping the similar images to clusters in content, it is applicable to attach additional textual metadata to images in the same clusters. This approach is helpful if there are no resources available to provide sufficient textual information. Nonetheless, Due to the intrinsic drawbacks of CBIR systems, this method can not explore the underlying semantics expressively.

Besides, all the related works mentioned above are only for a small dataset. Thus, the results are not quite persuasive and further research works are needed.

In this paper, we present a system, Illustrator, attempting to accomplish the illustration retrieval effectively and efficiently in a large illustration dataset. In our system, a semantic model is exploited to analyze the semantics of illustration. In this model, we first extract the shape information whose similarities are then used to construct two link graphs based on which we execute the auto-captioning procedure on the uncaptioned illustrations. In addition, cross-modal correlation analysis is applied to get rid of the noisy information and reduce the dimensionality of the feature vectors. Finally, we introduce a re-rank scheme that returns as many subtopics related to the query as possible along with the improvement in relevance. Experiments shows that our system performs efficiently on illustration retrieval and the illustrations in our experiments are all from digital books in the China-American Digital Academic Library (CADAL) project. To the best of our knowledge, Illustrator is the first system for retrieval in a very large illustration dataset.

The paper is organized as follows. Section 2 presents the architecture and basic techniques of text-based illustration retrieval systems, which is also used as the baseline in our experiments. We present our semantic-based system, Illustrator in Section 3. In Section 4, experiments and results are presented. Finally, we conclude our paper Section 5.

II. TEXT-BASED RETRIEVAL

The idea of text-based retrieval is that exacting the textual features as the main description to the illustrations. A typical architecture of text-based retrieval system is shown in Figure.1.

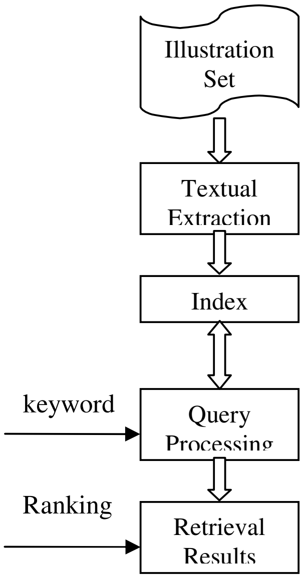


Fig. 1. Framework of text-based Retrieval system

Textual features are extracted from an Illustration Set (IS) and then indexed. Given a keyword by the user, the system executes the query on the index of the IS and returns the retrieval results by a certain ranking scheme

A. Textual Feature Extraction

All the digital books are subject to the OEB standard and pages are stored as HTML files. Since we only concerned with the illustrations in the books, pages without illustrations are eliminated in advance. From the HTML file, we extract the context in which an illustration appears. It is conspicuous that texts in different fields and positions have various effects on the illustrations. Thus, different weights should be assigned to the words according to their importance to the illustrations.

To begin with, illustrations are stored in JPG format, so the fields of the *img* tag are significantly important in describing the contents of the illustrations. The captions of the illustrations, if there is any, often appear in these fields. Therefore, words in these fields should be assigned to high weights. Besides, the fields of the *title* along with different fonts like *bold*, *italics* also contain rich information of the illustrations.

In addition, words around the illustrations, namely collateral texts, also play an important role in indicating the contents of the illustrations. Apparently, words appearing at different positions should be assigned with different weights due to their importance to the illustrations. We take the assumption that only 20 words before or after the illustration can help explaining the contents of illustrations. The weights assigned to the collateral words increases as the distances to the illustrations decrease.

B. Query

After extracting the textual features (e.g. captions, titles, collateral words), indexes are established based on the extracted textual features for further queries. Given a keyword by the user, the system executes the retrieval on the indexes and finds the relevant textual information, which is then used for mapping the corresponding illustrations. The whole procedure is shown in Figure 2.

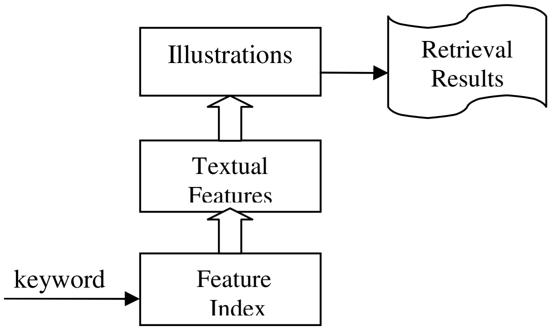


Fig. 2. The query procedure

When the dataset grows really large, the query procedure may be quite time-consuming. Therefore, the algorithm for query should be efficient and easy to scale to large dataset.

C. Ranking

In the query procedure, we have obtained the retrieval results which have various relations to the keywords. How to present the illustrations with the highest relevance to the users becomes a problem. We apply a ranking scheme in which each of the illustrations in the retrieval set is assigned to a certain score according to their relevance to the query.

As discussed in section 2.1, words in the fields of img, with the font like bold, italic as well as the collateral words around the illustrations contain semantic description of the illustrations. The ranking scheme is based the factors.

In [2], Marco La Casica *et al* proposed a scheme to assign different score to the texts around the illustrations. Words appear in the fields of *img* and in the font of *bold*, *italic* are assigned 5.0 and 4.0 respectively. Weights of the words around the illustrations depend on their distance of the illustrations. Taking the assumption that only 20 words before or after the illustrations count, we can computer the score as $5.0 * e^{-2.0 * pos/20}$, where pos is the position of word with respect to the illustration.

After combining all the scores together, we obtain the total score which can be considered to evaluate the relevance of the illustration to the query. Therefore, after ranking the scores, we can present the most relevant results to the users.

III. Semantic-based Retrieval

The traditional text-based retrieval we discussed above has its own drawbacks due to the limitations of the textual description of the illustrations. Therefore, visual features of the illustrations are exacted and a cross-modal analysis is applied to improve the semantic representation of the illustrations, namely semantic-based retrieval. The architecture of the retrieval scheme is shown in Figure.3.

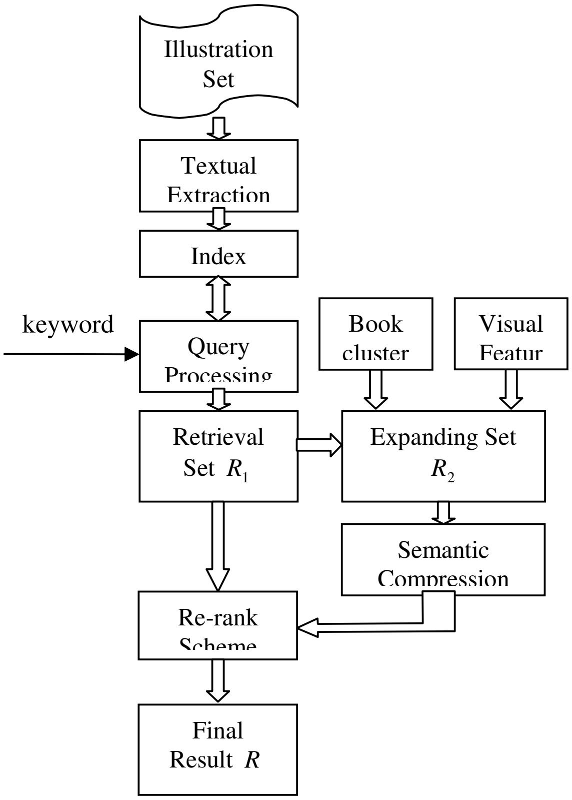


Fig. 3. Framework of semantic-based retrieval system

Given a keyword, we first apply the text-based retrieval engine to obtain a set in which the illustrations have textual relevance to the keyword. We denote the set as R_1 (see Figure.3). Then according to the similarities of the visual features, we group the illustrations into clusters which are used to expand the set we obtained. After applying some semantic compression techniques and a multi-rank scheme, we receive the ultimate retrieve results.

A. Textual Feature Extraction

We discussed the textual feature extraction in section 2. Nevertheless, the textual features are inadequate to describe the illustrations. Hence we extract the visual feature and exploit the underlying relations between the two kinds of features so that we can better represent the illustrations. There are mainly two forms of illustrations in the digital books. We denote the set I_1 as the illustration subset including charts, figures, tables, etc, which are mainly composed of graphics primitives (lines, marks, circular arcs, etc.). Among the numerous modalities (color, texture, shape, etc.) of pictorial data that can help indexing the illustrations, we focus the shape information for the indexing task. The other set I_2 includes the pictures of people, animals, landscapes, etc, which are mainly gray-scale images instead of the colored ones. Due to the different features of the two kinds of illustrations, we apply two methods to represent their visual features.

In [3], Cohen and Guibas proposed a method to compute the shape information for indexing. Based on the assumption that every illustration can be represented by the basic shapes through translation, rotation and scaling, they start with a collection of basic shapes $S = \{S_1, S_2, S_3, \cdots, S_k\}$. The elements in the set S are either built-in or user-defined types. For every illustration P, the index $\iota(P)$ of P is a subset of the set S, that is, a subsequence of $S = \{S_1, S_2, S_3, \cdots, S_k\}$ is extracted to match well into the illustration P. For each matching basic shape S_i, four parameters corresponding to the matches S_i into P are recorded. They are x_i and y_i (the translation), θ_i (the rotation) and $\log s_i$ (logarithm of the scale). We denote $T_i(x_i, y_i, \theta_i, s_i)$ as the matching transformation of S_1 into P.

The images of people, animals, landscapes, etc are in various and complicated shapes, therefore, shape features we discussed in section 3.1.1 can not represent their visual features properly. The most widely used visual features include: (1) color features such as color histogram, color moment, color coherence vector; (2) texture features such as edge histogram, co-occurrence matrix and Gabor wavelet features; (3) shape feature such Fourier descriptor and moment invariant. We use moment invariants to measure shape information of the illustrations. We compute a seven dimensional feature vector of seven moments and each moment is invariant to translation, rotation and scaling. Although the features we extracted can not perfectly represent the original illustrations, they could provide correlations in visual features among illustrations and therefore supply us an alternate method to dig the underlying semantic correlations among illustrations.

B. Retrieval Set Expansion

The illustrations retrieved in R_1 are textual-relevant to the keyword. However, the text features are insufficient to express the semantics of the illustrations. Therefore, we exploit the similarities of visual features to expand the set R_1.

In order to expand the set, we construct two graphs with the names of BSSF graph and MIF graph based on the basic shape features and moment invariant features respectively. Figure 4 depicts the main idea of the set expansion.

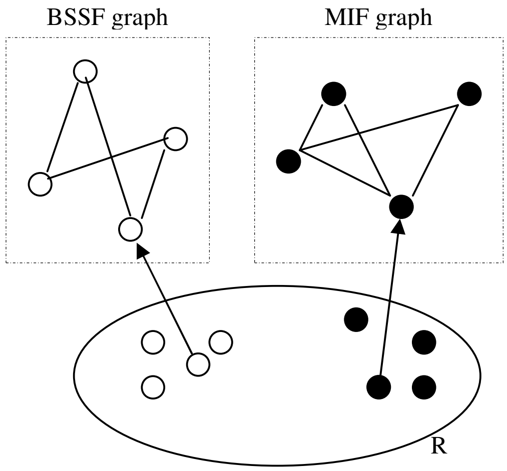


Fig.4. Retrieval set expansion

To construct the BSSF graph, given two illustrations $P, Q \in I_1$, Let $\imath(P)$ and $\imath(P)$ denote indexes of P and Q, $T_P = \{T_P^1, T_P^2, \cdots, T_P^p\}$ and $T_Q\{\{T_Q^1, T_Q^2, \cdots, T_Q^q\}\}$ denote the matching transformation of P and Q, where $T_P^i(x_P^i, y_P^i, \theta_P^i, s_P^i)$ $(1 \le i \le p)$ denotes the matching transformation of S_P^i into P, $T_Q^j(x_Q^j, y_Q^j, \theta_Q^j, s_Q^j)$ $(1 \le j \le q)$ denotes the matching transformation of S_Q^j into Q, we use the Hausdorff distance to measure their similarity.

$$sim(T_P, T_Q) = H(T_P, T_Q) = \max(h(T_P, T_Q), h(T_Q, T_P)) \quad (1)$$

Where

$$h(T_P, T_Q) = \sup_{\alpha \in T_P} \inf_{\beta \in T_Q} d(\alpha, \beta) \qquad (2)$$

$d(\cdot,\cdot)$ is the distance between two points in vector space. A link from P to Q with weight $sim(P,Q)$ is constructed if $sim(P,Q) \ge \varepsilon$ (ε is a threshold); otherwise no link in constructed. What is worth being noted is that the BSSF graph can only be applied among illustration belonging in I_1.

For the illustrations in I_2, we construct the MIF graph. Each illustration can be represented as a vector $\vec{v}_i$. We simply calculate the similarity between $\vec{v}_i$ and $\vec{v}_j$ as

$$sim(\vec{v}_i, \vec{v}_j) = \cos(\vec{v}_i, \vec{v}_i) = \frac{\vec{v}_i \cdot \vec{v}_j}{\| \vec{v}_i \| \cdot \| \vec{v}_j \|} \qquad (3)$$

Again, a threshold δ is set and we note that the MIF graph can only be applied among illustrations in I_2.

Thus, each link in the two graphs has been assigned a weight to indicate the similarity between the corresponding illustration pair. Usually, illustrations expressing the same meaning are similar in their visual features to each other. Hence, in both graphs, a group of heavily linked documents potentially represent a semantic group, illustrations connected by weak or no links contain different semantics. Based on the two graphs, we expand R_1 and obtain R_2.

C. Cross-Modal Correlation (CMC) Analysis

We exploited the textual information to aid the retrieval of illustrations. However, there are approximately one third illustrations are uncaptioned or the textual features are usually companied with some noisy and irrelevant information. Our task is to accomplish the auto-captioning, which is defined as follows:

Auto-Caption. *Given a set of illustrations, each is with a visual description by a 7-D vector. Some of them are captioned by several keywords, which some of them lack the textual information. Find the best words that describe the uncaptioned illustrations.*

In [9], Pan *et al* proposed a method that based on the image segmentation and random walk techniques. However, the method can not be applied directly to our problem owing to the difficulties to segment the illustrations. To solve the problem, we assume that illustrations with similar visual features probably have potential textual similarities and our ideas are shown in Figure 5.

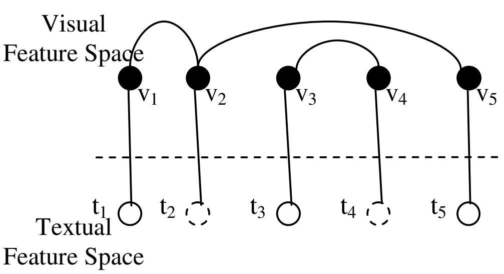


Fig.5. Auto-captioning

As illustrated in Figure 5, v_1, v_2, v_3, v_4, v_5 are visual features of 5 illustrations. The circles composed by solid lines t_1, t_3, t_5 are textual features corresponding to v_1, v_3, v_5, while there are no corresponding textual features to v_2, v_4. To solve the problem, we generate the virtual textual features (the circles composed by dashes) by analyzing the linkages among visual features as well as the correlations between visual features and textual features. We obtain the visual feature linkage formations in the link graph (BSSF graph or MIF graph). In Figure 5, v_1, v_5 have visual links with v_2, therefore, we can obtain the textual information t_2 through t_1 and t_5. In the same way, we receive t_4 through t_3 and t_5. In general, given an uncaptioned illustration with the visual feature v_i, the corresponding textual features t_i is calculated as

$$t_i = \sum_{v_j \in D} \omega_{ij} \cdot t_j \qquad (4)$$

Where $D_i = \{v_j \mid (v_i, v_j) \in E\}$, E is the edge set of the link graph, w_{ij} is the weight between v_i and v_j.

The visual and textual features can be combined into high-dimensional vectors and then be directly used for retrieval. Let *n* stand for the dimensions of visual features, and *m* stand for the dimensions for textual features, and then in the joint visual-textual features space, each illustration can be represented as

$$\begin{aligned} f_i^c &= [f_i^V, f_i^T] \\ &= [v_{i,1}, \cdots, v_{i,j}, \cdots, v_{i,n}, t_{i,1}, \cdots, t_{i,k}, \cdots, t_{i,m}] \end{aligned} \qquad (5)$$

Note that since various visual and textual features can have quite different variations, we also need to normalize each feature in the joint space according to its maximum elements (or certain other statistical measurements).

However, the dimensionality of the visual or textual feature vectors is pretty high. For reducing the feature dimensionality, an often-used method is the so-called latent semantic indexing (LSI) technique [5], which relies on singular value decomposition (SVD) of the feature matrix to capture the latent semantic structure among the matrix elements. Thus we can apply the LSI technique to reveal the latent semantic structure in the joint visual-textual feature space, as in [12]. However, LSI does not distinguish features from different modalities in the joint space, thus the optimal solution based on overall distribution may not best represent semantic relationships between features of different modalities. In [8], two cross-modal association analysis methods, i.e., cross-modal factor analysis (CFA) and canonical correlation analysis (CCA), were introduced to identify and measure intrinsic associations between visual and audio features. Here we adopt the CFA method to capture the best coupled patterns between visual and textual features.

The key idea underlying the CFA method is to find two orthogonal transformation matrices so that the coupled data in the two subsets of features can be projected as close to each other as possible [8]. Let $\boldsymbol{F}_\mathrm{v}$ and $\boldsymbol{F}_\mathrm{T}$ be the visual and textual feature matrices for the illustrations in $\boldsymbol{R}_2$, then the transformation matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ can be obtained by solving the following optimization:

$$\min \| F_V A = F_T B \|_F^2 \quad s.t. \quad A^{'} A = I, B^{'} B = I \qquad (6)$$

Where $\|\cdot\|_F$ denotes Frobenius norm. According to the orthogonality of $\boldsymbol{A}$ and $\boldsymbol{B}$, we have

$$\|F_V A - F_T B\|_F^2 = tr(F_V F_V') + tr(F_T F_T') - 2tr(F_V A B' F_T') \qquad (7)$$

Where $tr(\cdot)$ denotes the matrix trace. Thus Eq.(6) is equivalent to maximize the term $2tr(\tilde{V} A B' \tilde{T}^T)$. It can be shown [8] that such matrices are given by the SVD decomposition of $F_V' F_T$, i.e., $F_V' F_T = ADB$, where $\boldsymbol{D}$ is the singular value matrix. Thus with the optimal transformation matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, F_V and F_T can be transformed by the following equation:

$$\begin{cases} \tilde{F}_V = F_V A \\ \tilde{F}_T = F_T B \end{cases} \qquad (8)$$

$\tilde{F}_V$ and $\tilde{F}_T$ can then be combined into a joint feature matrix $\tilde{F}_C$. Similarly to those in LSI, the first and most important *k* vectors in $\tilde{F}_V$ and $\tilde{F}_T$ can be used to preserve the principal coupled patterns in much lower dimensions.

A significant advantage of the CFA method is in favor of coupled patterns with high variations. However, the CFA method is based on some naïve techniques such as the linear correlation model and the projected distance, which would limit its applications in more complex situations. Moreover, such an approach can only be applied in the cases where the feature matrix for all the testing illustrations is constructed offline. Instead, in this paper two optimal transformation matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are learned from the training illustrations in $I_{(\mathrm{B})}$, and then used as two semantic matrices to map the testing illustrations in $I_{(\mathrm{T})}$ into another semantic space. That is, for a target illustration $P_i \in I_{(T)}$, the transformed feature vector can be represented as $f_i^c = [\tilde{f}_i^V, \tilde{f}_i^T]$, where

$\tilde{f}_i^V = f_i^V A$ and $\tilde{f}_i^T = f_i^T B$. This is similar to the semantic smoothing method used in [4][10]. Without risk of confusion, this paper refers to this version of the CFA method as cross-modal correlation (CMC) analysis.

Thus if we consider the visual-textual joint space, the corresponding kernel is given by $K_C = [k(f_i^C, f_j^C)]$, where $k(f_i^C, f_j^C)$ is a kernel function. Here we refer to K_C as the content kernel.

D. Re-rank Method

Most current retrieval systems intend to provide the retrieval results to users according to the relevance scores of each result to the query. The idea is quite eloquent when the users are clear about what they need and provide the exact key to the system. However, in most occasions, the users fail to present the accurate keywords and therefore their intentions are expressed ambiguously.

It is a good idea that the system returns as many subtopics of the query topic as possible without the loss of relevance. In this case, users can pick up the illustrations they are interested in and therefore the possibility that users get satisfactory results enhances significantly. However, in traditional retrieval research, precision and recall are two factors to evaluate the performance of retrieval system. But both of the two criteria only concern the relevance of the results, without considering the variety of the topics the results cover. Therefore, the top search results are often bounded into a group of closely related topics and fail to meet the needs of the users' diversified requirements.

To overcome such situation, we introduce a re-rank scheme based on a penalty algorithm. Given the link graphs we constructed in section 3.2, the scores we computed in section 2.3 and the expanded retrieval set R_2, the penalty algorithm are executed as follows:

```
Step 0. Initialize the score $s_i$ of each illustration $P_i$ in set
        $R_2$, the final result set $R = \Phi$
Step1.  Sort the illustrations in $R_2$ by their scores in
        descending order.
Step2.  Suppose the illustration ranked highest in $R_2$
        is $P_i$. Add $P_i$ into $R$ and delete $P_i$ in $R_2$, then
        find all the illustrations $P_j$ that have a link to $P_i$
        in the link graphs and impose a penalty to the
        score of each $P_j$ as follows:
                $s_j = s_j - sim\left(f_i^c, f_j^c\right) \cdot s_i$
Step3.  Update the scores and re-sort the illustrations in
        $R_2$ in descending order.
Step4.  Go to Step 2 until $B = \Phi$
```

Note that the vital part of the above algorithm is Step2, which contains the penalty idea that reduce the scores of illustrations which have semantic link with the chosen illustrations in the final result set. By executing the penalty algorithm, we always keep the most informative illustrations in the top of R_2 and therefore provide a various subtopics in the final result set.

IV. EXPERIMENTS

We conducted experiments to demonstrate the retrieval result of the Illustrator system outperform that of the traditionally text-based retrieval systems. All the illustrations along with the textual pages used in our experiments are obtained in digital books. The dataset contains approximately 500,000 illustrations, we filtered the illustrations whose width and height are both smaller than 100 pixels, and whose ratio between width and height are greater than 5 or less than 1/5.

A. Relevance

The recall (R) and precision (P) are often used to evaluate the relevance of the results to the query. Due to the huge number of illustrations in the dataset, it is quite difficult to label all the results in advance. Hence only the precision is considered as the criterion. In addition, we only concerned the top 20 results. We picked up 20 keywords and executed the query. Figure 6 shows that Illustrator significantly improves the value of precision compared to the text-based retrieval (TBR) systems.

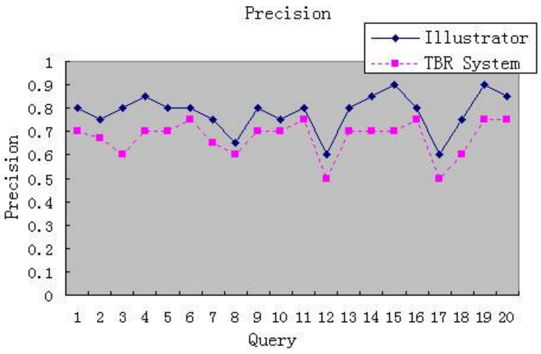


Fig. 6. Precision of Illustrator and text-based retrieval system

B. Diversity

The factor diversity is used to evaluate the number of topics related to the query in the returned illustrations. Again, we picked up 20 keywords and executed the query and only the top 10 results are taken into consideration. Figure 7 shows that the Illustrator system provides much more topics than the text-based retrieval (TBR) systems.

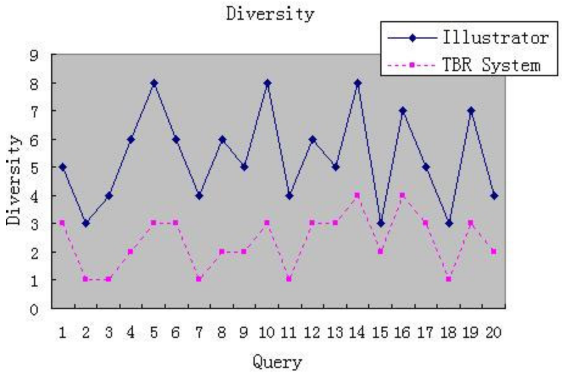


Fig. 6. Diversity of Illustrator and text-based retrieval system

C. Dimension Reduction

We also performed experiments to evaluate the effects of the LSI and CMC analysis on our dataset. When illustrations are represented by textual features, we can exploit LSI analysis to reduce the feature dimensionality; when illustrations are represented by both visual and textual features, both LSI and CMC can be used.

Table 1 shows that average dimension reduction ratios (DR). We can see that both LSI and CMC can effectively reduce the dimensionality of the feature space. Comparatively, the CMC analysis has large DR. Note that here we use the same eigengap based method to determine an appropriate *k* value for LSI and CMC.

TABLE I
THE COMPARISON OF AVERAGE DIMENSIONAL REDUCTION RATIOS

	LSI (textual)	**LSI (visual-textual)**	**CMC (visual-textual)**
Before	1036	1043	1043
After	312	315	213
DR	3.32	3.31	4.90

D. Efficiency

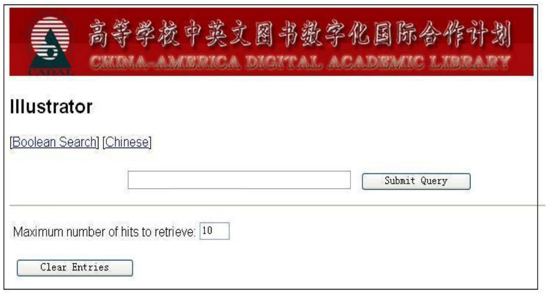
We picked up 50 keywords and executed the queries. by adjusting the number of the illustrations returned, we change the conditions for retrieval so that the results are more comprehensive and persuasive. From Table 2, we can tell that, compared with the baseline, the semantic model we proposed almost has no effect in lowering the performance of the retrieval system on various conditions.

TABLE II
AVERAGE QUERY TIME ON 50 KEYWORDS

	10 results	**20 results**	**50 results**
Illustrator	3.4s	3.5s	3.5s
Baseline	3.2s	3.3s	3.2s

E. A Case Study

We provide a case study to show how our system works. Figure 7 shows the interface for Illustrator. We input the word "forest" as the query keyword and the top 5 results are shown in Figure 8.



高等学校中英文图书数字化国际合作计划
CHINA-AMERICA DIGITAL ACADEMIC LIBRARY

Illustrator

[Boolean Search] [Chinese]

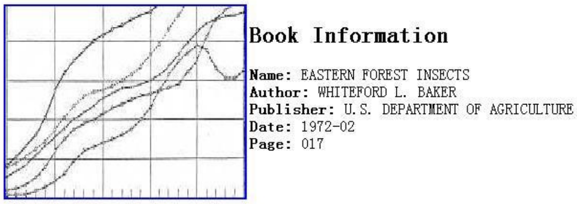
Submit Query

Maximum number of hits to retrieve: 10

Clear Entries

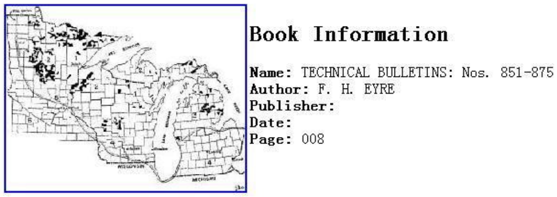
Fig. 7. System interface

The retrieval results contain four subtopics of forest. They are forest insect, forest coverage in various districts, forest overview and endangered forests. Book information are listed for users' convenience.



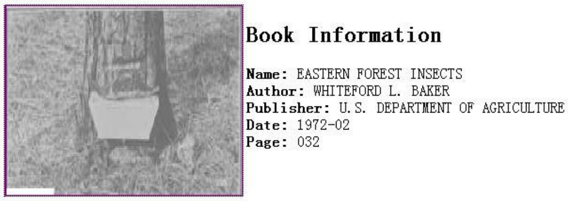
Book Information

Name: EASTERN FOREST INSECTS
Author: WHITEFORD L. BAKER
Publisher: U.S. DEPARTMENT OF AGRICULTURE
Date: 1972-02
Page: 017



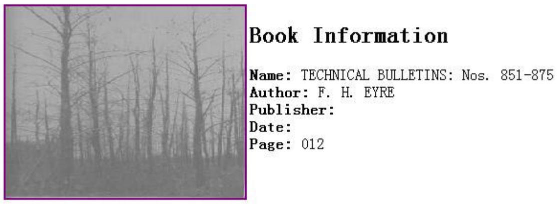
Book Information

Name: TECHNICAL BULLETINS: Nos. 851-875
Author: F. H. EYRE
Publisher:
Date:
Page: 008



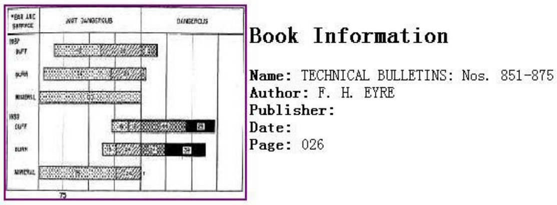
Book Information

Name: EASTERN FOREST INSECTS
Author: WHITEFORD L. BAKER
Publisher: U.S. DEPARTMENT OF AGRICULTURE
Date: 1972-02
Page: 032



Book Information

Name: TECHNICAL BULLETINS: Nos. 851-875
Author: F. H. EYRE
Publisher:
Date:
Page: 012



Book Information

Name: TECHNICAL BULLETINS: Nos. 851-875
Author: F. H. EYRE
Publisher:
Date:
Page: 026

Fig. 7. Retrieval results of the query word “forest”

V. Conclusion

In this paper, we present a semantic-based illustration retrieval system, Illustrator. The main objective of Illustrator system is to exploit the visual and textual information to aid the retrieval of illustrations and overcome the drawbacks of the traditional text-based retrieval systems. Our main contributions are summarized as follows:

Due to the drawbacks of the architecture of text-based retrieval system, we introduced a novel scheme of illustration retrieval and developed a retrieval system, Illustrator.

In addition, we proposed a CMC model which is exploited to capture the correlation among different modals of features of illustrations.

Moreover, we introduced a re-rank scheme which diversifies the topics in retrieval results significantly.

The Experiment results demonstrate the effectiveness of Illustrator system in illustration retrieval

Acknowledgement

The project in our paper is supported by China-American Digital Academic Library (CADAL) project (No. CADAL2004002) and the Hi-Tech Research and Development Program (863) of China (No. 2004AA119010)

References

[1] Borowski, M., Bröcker, L., Heisterkamp, S., Löffler, J. Structuring the Visual Content of Digital Libraries using CBIR Systems, *In proceedings of IEEE International Conference on Information,* 288-293, 2000

[2] Cascia, M.L., Sethi, S. and Sclaroff, S., Combining Textual and Visual Cues for Content-Based Image Retrieval on the World Wide Web, In *Proceedings of the IEEE Workshop on Content - Based Access of Image and Video Libraries*, 1998

[3] Cohen, S.D. and Guibas, L.J., Shape-based Illustration Indexing and Retrieval - Some First Steps, In *Proceedings of the ARPA Image Understanding Workshop, 1209-1212, February 1996.*

[4] Cristianini, N., Shawe-Talyor, J., Lodhi, H., Latent semantic kernels. *Journal of Intelligent Information Systems*, **18**(2/3):127-152, 2002.

[5] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., Indexing by latent semantic analysis. *Journal of American Society of Infomation Science and Technology*, **41**(6):389-401 1990.

[6] Hu, Y.H., Li, H., Cao, Y.B., Meyerzon, D and Zheng, Q.H., Automatic Extraction of Title from General Documents using Machine Learning, In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries,* 145-154, 2005

[7] Hu, Y.H., Xin, G.M., Song, R.H., Hu, G.P., Shi, S.M., Cao, Y.B., and Li, H. Title extraction from bodies of HTML documents and its application to web page retrieval, In *proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 250-257, 2005.

[8] Li, D.G., Dimitrova, N., Li, M.K., Sethi, I.K., Multimedia Content Processing through Cross-modal Association. Proceedings of 11th ACM International Conference on Multimedia, Berkeley, California, USA, p.604-611, 2003.

[9] Pan, J.Y., Yang, H.J., Faloutsos, C. and Duygulu, P., Automatic multimedia cross-modal correlation discovery, In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 653-658, 2004.

[10] Siolas, G., d’Alché-Buc, F., Support Machine Learning Based on Semantic Kernel for Text Categorization. Proceedings of the International Joint Conference on Neural Network (IJCNN), 2000.

[11] Tian, Y.H., Huang, T.J., Gao, W., Exploiting multi-context analysis in semantic image classification. *Journal of Zhejiang University SCIENCE*, Vol.6A,No.11,pp1268-1183, 2005.

[12] Zhao, R., Grosky, W.I., Narrowing the semantic gap — improved text-based Web document retrieval using visual features. *IEEE Trans. Multimedia*, 4(2):189-200, 2002.