Metadata Extraction and Indexing for Map Search in Web Documents

Qingzhao Tan
Department of Computer Science and Engineering
Pennsylvania State University
University Park
PA 16802, USA
qtan@cse.psu.edu

Prasenjit Mitra
College of Information Sciences and Technology
Pennsylvania State University
University Park
PA 16802, USA
pmitra@ist.psu.edu

C. Lee Giles
College of Information Sciences and Technology
Pennsylvania State University
University Park
PA 16802, USA
giles@ist.psu.edu

ABSTRACT

In academic scientific articles, maps are widely used to provide the related geographic information and to give readers a visual understanding of the document content. As more digital documents containing maps become accessible on the Web, there is a growing demand for a Web search system to provide users with tools to retrieve documents based on the information available within a document's maps. In this paper, we design methods and algorithms to extract, identify, and index maps from academic and scientific documents in digital libraries. Experimental results show that our approach can accurately locate maps and significantly improve the retrieve quality for maps in digital documents.

Categories and Subject Descriptors: H.4[Information Systems Applications]: Miscellaneous
General Terms: Algorithms, design, experimentation
Keywords: Digital library, extraction, indexing, ranking function.

1. INTRODUCTION

Maps are a type of important figures which are frequently used in digital documents, especially academic and scientific documents in the field of archeology and geography. Despite the importance of conveying information using maps, to our best knowledge, no digital library has paid attention to geographical information in maps that are contained in digital documents. In this paper, we address this problem by integrating the map search functionality into a scientific digital library.

To build a map search system, we first construct a set of metadata for each map in a digital document. This metadata includes map level metadata (e.g., a map's caption) as well as document level matadata (e.g., a document's title, abstract). We denote the document containing the map the *host document* of that map. We then provide methods and algorithms to achieve the following three tasks: 1) *Map Metadata Extraction*. A set of map metadata are extracted from the digital document. We first use TET [2] to extract text from PDF files and then apply a set of heuristics rules to locate the metadata in the extracted text. 2)*Map Identification*. Maps are identified from the other non-map images using supervised learning methods and a set of features which are extracted from the map metadata. 3) *Map Indexing and Retrieval*. Regarding the map metadata as various fields for map indexing, a novel index and a new ranking function are developed for the map search system.

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

2. MAP SEARCH SYSTEM

A document is essentially a combination of images and text. Understanding the relationship between an image and its accompanying text in the host document can reveal valuable information for interpreting the image. We first define a set of metadata which come from two sources: the map level metadata which is generated from the text accompanying the map; the host document level metadata which is generated from the host document's metadata. On one hand, within a host document, there are three resources which bear metadata for the map content: the caption, the reference text, and the document page containing the map. On the other hand, from the map's host document, we extract the document level matadata, which are composed of *title*, *abstract*, *author*, *age (i.e., publication date)*, and *citations*.

We then consider the map identification process as a classification process, which is to categorize the images into two groups, maps and non-maps. The map identifier is a supervised classifier which works in two phases: training and testing. First, a classifier is trained on the feature vectors generated from the training set in which each image has a class label indicating whether it is a map or not. This classifier is then used to test other images. Cross validation [4] is used for the training and the testing process. Considering that our problem is in some degree similar to the text classification problem, we propose using Support Vector Machines (SVMs), which is a learning machine for two-group classification problems [5].

Finally, we define a novel scheme named weighted Map Term Frequency and Inverse Map Term Frequency (*MTF-IMTF*) for our map search system. Given a map M_j, for a term T_i which appears in the kth field MD_k with W_k, we define the weighted *MTF* as $W_k MTF_{ijk}$. Here MTF_{ijk} is the map term frequency of the term T_i in the metadata MD_k of map M_j. For *IMTF*, we use N_m to denote the total number of maps in a collection, and use n_{ik} to denote the number of maps which contain the term T_i in its metadata MD_k. Therefore, ω_{ijk} is computed as, $\omega_{ijk} = W_k \times MTF_{ijk} \times \log N_m/n_{ik}$. Similarly, we can get ω_{iqk} for a query Q if Q specifies the metadata to be searched for. Otherwise, if Q consists of only a set of keywords, it is supposed that each keyword appears in each MD. After computing the sets of ω_{ijk} and ω_{iqk}, the map M_j and the query Q are represented as D-dimensional vectors $\overrightarrow{m}_j$ and $\overrightarrow{q'}$, respectively. Here $D = \sum_{i=1}^{f} |MD_i|$. Finally, the similarity between a map M_j and a query Q is computed as the cosine of the angle between $\overrightarrow{m}_j$ and $\overrightarrow{q'}$ as, $sim(M_j, Q) = \frac{\overrightarrow{m}_j \bullet \overrightarrow{q'}}{|\overrightarrow{m}_j| \times |\overrightarrow{q'}|} = \frac{\sum_{i=1}^{D} \omega_{ijk} \cdot \omega_{iqk}}{\sqrt{\sum_{i=1}^{D} \omega_{ijk}^2} \sqrt{\sum_{i=1}^{D} \omega_{iqk}^2}}$.

# of Papers	2000
# of Papers with Fig.	465
# of Fig.	2090
# of Papers with Maps	278
# of Maps	536

Table 1: Statistics of the collection used in the tests.

Approaches	$P(\%)$	$R(\%)$
Begins with "map"	100	23.1
$\bigcup(AllRules)$	37.2	92.7
SVM	88.7	91.6

Table 2: Map identification performance. SVM outperforms the rule-based methods.

We currently customize our $MTF - IMTF$ scheme with five sets of metadata: caption, reference text, document title, document abstract, and a set of location names which are extracted from all the above. The weights of these five sets of metadata are denoted as W_c, W_r, W_t, W_a, and W_{loc}, respectively. By setting and tuning the weights, we obtain various ranking strategies.

Besides the relevance of the map's content, we use several query-independent features of the map and its host document to adjust the final ranking. We denote these two sets of features as MF and HDF, respectively. In MF, we take into account two factors: 1) the length of map caption L_c; and 2) the length of map reference L_r. Boost by MF, the maps with more detailed explanations get higher ranking. In HDF, we take into account four factors: 1) the map frequency $MFreq$; 2) the number of other publications citing the host document $Cite$, which can be obtained from Google Scholar [1]; 3) the venue's prestige VP, which can be obtained from some online journal ranking website; and 4) the document freshness DF, which is indicated by the publication year. Boost by HDF, those maps whose host documents contain more maps, newly published in high quality journals/references, and are cited by more publications can get higher rankings. Finally, normalizing all the above six features, we get the ranking score for M_j as, $score(M_j, Q) = sim(M_j, Q) \times MF \times HDF$, where $MF = L_c + L_r$, $HDF = MFreq + Cite + VP + DF$.

3. EXPERIMENTAL EVALUATION

Experiments were carried out to verify the effectiveness of our proposed map search system. The collection we used to perform experiments consists of a set of archeology documents, which are downloaded from the scholarly journal archive, JSTOR (http://www.jstor.org/). We manually analyzed each document and listed some statistics of the collection in Table 1.

Metadata Extraction and Map Identification To evaluate the performance of metadata extraction and map identification, we use $precision(P)$ and $recall(R)$ as the measures.

We adopted the well-known SVM-Light (http://svmlinght.joachims.org/) for our training and testing on the 2090 figures. Default parameter setting, i.e., linear kernel, is used. We compared the SVM method with two rule-based methods: 1) use only one feature, BeginsWith "map", which can get the highest P, 100%, but very low R, 23.1%; 2) use the union of all positive features, which can get the highest R, 92.7%, but very low P, 37.2%. Table 2 lists the performance evaluation using five-fold cross validation. As expected, the comparison results illustrate that the SVM method shows significant improvement to both the two rule-based methods.

Map Ranking We used 25 keyword-based queries to search in the 536 maps which are contained in 278 papers. The average query length is 1.32 words. A pool method was used to determine the set of relevant maps to each test query. We ran our ranking methods for each query and put the 30 highest ranked maps in each returned ranking into a pool for evaluation. As a result, for each query, we have a set of maps, labeled as relevant or non-relevant, independently of the ranking functions. We used $P - R$ figures [3] to demonstrate the quality of ranking results.

Eight ranking variances are implemented and compared using *P-R* figures: 1) customized Google Desktop search engine, which is a Google Desktop API customized to build up index on the directory containing the 278 PDF files; 2) full text indexing; 3) caption only, which means $W_c = 1, W_r = W_t = W_a = W_{loc} = 0, MF = HDF = 1$; 4) reference only, which means $W_r = 1, W_c = W_t = W_a = W_{loc} = 0, MF = HDF = 1$; 5) Caption+Reference, which means $W_c = W_r = 0.5, W_t = W_a = W_{loc} = 0, MF = HDF = 1$; 6) Caption+Reference+DocMeta, which means $W_c = W_r = W_t = 0.33, W_a = W_{loc} = 0, MF = HDF = 1$; 7) weighted Caption+Reference+DocMeta+Locations, which means $W_c = W_r = W_{loc} = 0.25, W_t = W_a = 0.125, MF = HDF = 1$; 8)weighted and boost Caption+Reference+DocMeta+Locations, which means $W_c = W_r = W_{loc} = 0.25, W_t = W_a = 0.125, MF = L_c + L_r, HDF = MFreq + Cite + VP + DF$. For 1) and 2), we tested the queries with an additional keyword, "map". Their returned results are not maps but documents. We located the page of the "high lighted text" in each document, which is a text segmentation considered to be the most relevant to the test query by the ranking strategy. Then the map on that page is considered to be the result. If there is no map on that page, a miss is counted.

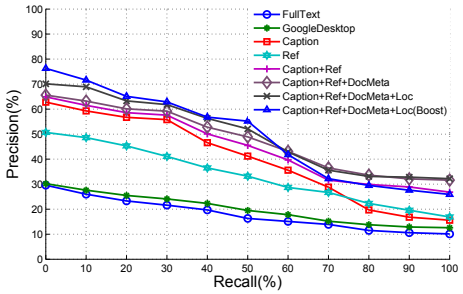

Figure 1: The average precision-recall figures for 25 queries.

The average *P-R* figures for the 25 test queries are shown in Figure 1. From this figure, we can see that, 1) and 2) have the worst performance because they do not take any effort to deal with maps. 3) and 4) can both locate the maps in the documents and returned relevant maps to the query. 3) achieves higher precision than 4) before the 80% R because a map's caption concentrates on the map's content while the subjects of the references may be other content in the main body. By combining 3) and 4), 5) got higher precisions. This shows that indexing based on the combination of caption and reference can provide high quality results for maps. After including the document metadata, 6) presents a slightly better retrieval performance. This confirms that the document's title and abstract can give some useful information for the map indexing. Location keywords are included and given higher weights than document metadata in both 7) and 8). The *P-R* figures illustrate that the inclusion of location keywords leads to an increase in the average precision. Particularly, 8) got the highest average precision when average recall is below 60%. This means that the boost factor can greatly improve the quality of top ranked results. This is especially helpful to a digital library system, where precision is very important among the top ranked documents.

4. CONCLUSIONS

We have proposed a map search system for digital academic documents. Experimental results performed on JSTOR archaeology journal document show promising results.

5. REFERENCES

[1] Google scholar. http://scholar.google.com/.

[2] Pdflib tet - text extraction toolkit. http://www.pdflib.com/products/tet/.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[4] P. A. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach*. Prentice-Hall, London, 1982.

[5] T. Joachims. Making large-scale support vector machine learning practical. pages 169–184, 1999.