On Indexing Mechanism in Geographical Information Retrieval System

Xing LIN*, Bo YU**, Yifang Ban
*Geoinformatics, Royal Institute of Technology (KTH), Stockholm, Sweden
** Information Sciences and Technology, the Pennsylvania State University, USA

INTRODUCTION

With the development of Concerning the Geographical Information Retrieval (GIR) Systems (Larson, 1995), a lot of available indexes for GIR have been invented towards such problem. Zhou (Zhou, 2005) proposed a hybrid approach of inverted file and R-tree family. Vaid (Vaid, 2005) presented the geographic and full-text indexes in the famous GIR system – SPIRIT. Martins (Martins, 2005) studied nearly all the important indexing and ranking approaches in Information Retrieval (IR) systems and suggested a practical guide on how to build efficient a GIR system based on these approaches.

In the following part of this paper, a new hybrid indexing method is proposed using space-filling curve (SFC) and inverted file. Consequently, some experiments are carried out to compare this new index with the others. Finally, some conclusions are made according to the theoretic analysis and the result of experiments.

INDEXES IN GEOGRAPHICAL INFORMATION RETRIEVAL SYSTEM

Generally, the usual indexing mechanisms in GIR systems could be categorized into four groups: *Pure Keyword Index (PKI), Keyword-Spatial Double Index (KSDI), Spatial-Keyword Hybrid Index, (SKHI)* and *Keyword-Spatial Hybrid Index (KSHI).*

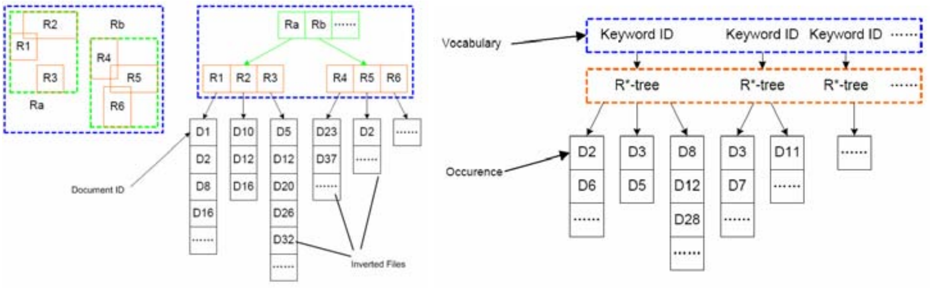

Figure 1. Two Hybrid indexes: SKHI (left) & KSHI (right).

A good indexing mechanism for GIR systems should take the 3 aspects into accounts: efficiency, storage overhead and operability. As stated above, a good indexing mechanism doesn't need to be prominent in some of the three aspects, but need to achieve a harmony among all these three aspects. From this point of view, among the four possible indexing mechanisms in GIR systems: (1) PKI is easy to implement but suffers from a low efficiency because of the sequent scanning in geographical scope; (2) KSDI will also enjoy an advantage of easy to implement and moderate efficiency, but it will taken up too much space to hold the two individual indexes. (3) Both SKHI and KSHI should win a higher efficiency in the process of information retrieval, but neither of them is good at reducing the storage overhead and avoiding ruining the existing keyword based indexing mechanisms.

A NEW HYBRID INDEX BY MERGING INVERTED FILE AND SPACE-FILLING CURVE (SFC)

Practically, our new hybrid index for GIR tries to use a certain type of space-filling curve to assign a linear label to each document according to their geographic scope. Then such labels will be inserted into the occurrence table of an inverted file as a spatial attribute for the documents. As a result, this new hybrid index shares nearly the same structure with normal inverted file. What is different is that each document in occurrence table will be referenced by two tags: *Document ID* and *Geographic ID*, as shown in the following pictures. Practically in order to speed up the spatial search, the corresponding occurrence table for each keyword could be rearranged according the geographic id of each document in an order of B-tree or alphabet.

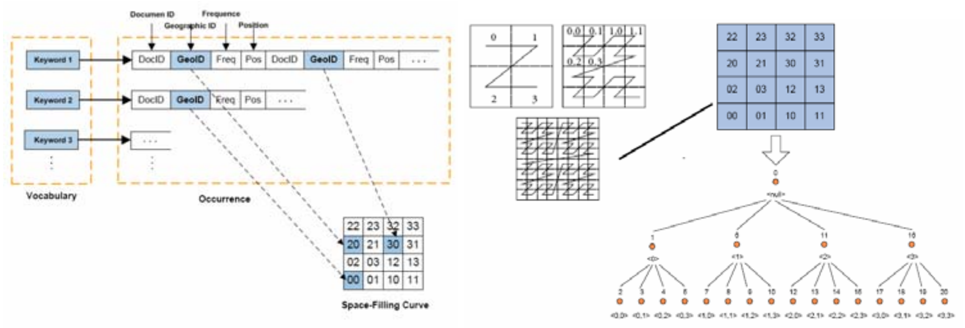

Figure 2. (a) Structure of the hybrid index based on SFC and inverted file. (b) Z-Filling Curve.

Obviously, this new hybrid index for GIR is quite easy to implement into the modern IR systems based on inverted files. Because structurally it only introduce a single geographical id to traditional inverted file, the same file organization and compression approaches of inverted file could be applied without any overhead here. Concerning the storage overhead of indexes, it will also be as small as a *O(N)* level.

EXPERIMENTS

The experiments are devised to evaluate the efficiency of this new approach proposed in the paper comparing a typical *KSDI* and another *KSHI* index. It is performed in a prototype system built using DotLucene (DotLucene, 2004) and Mg R-tree library (Pavlata, 2004). All the data and queries used here are generated in a random way and the results of evaluation are shown in the figures below (Fig.3).

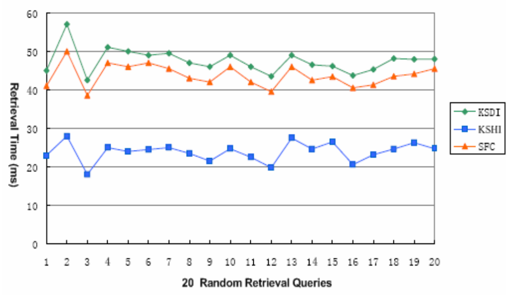

Figure 3. Evaluation of the efficiency of KSDI, KSHI and SFC.

Seen from the Fig.3, the SFC based hybrid index is not the fast one while comparing the KSDI (double indexes: R-tree index in geographic scope and inverted file index in thematic scope) and KSHI (hybrid index: inverted file in thematic scope and R-tree in geographic scope) which is the most efficient one. But in normal case the SFC based approach proposed here is always faster than the KSDI indexes.

CONCLUSION

In this paper, a new hybrid indexing mechanism based on space filling curve and inverted file is proposed for modern GIR systems. Although it is not the fastest one comparing to other possible ones, it is still of great potential for a real GIR application, especially for extending the information systems to be GIR competent because of its less storage overhead and little harm to the structure of existing inverted file.

BIBLIOGRAPHY

A. Guttman. A dynamic index structure for spatial searching. In Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 47-54, 1984.

Ray R. Larson. Geographic information retrieval and spatial browsing. In Linda C. Smith and Myke Gluck, editors, Proceedings of the Data Processing Clinic - Geographic Information Systems and Libraries: Patrons, Maps, and Spatial Information, Geographic Information Systems and Libraries: Patrons, Maps, and Spatial Information, 1995.

DotLucene. A powerful open-source search engine for .NET, 2004. Online at: http://www.dotlucene.net/.

Ondrej Pavlata. Mg R-tree Library: A simple r-tree implementation with c++ source code. 2004 Online at: http://www.volny.cz/r-tree/.

Yinghua Zhou, Xing Xie, ChuangWang, Yuchang Gong, andWei-Ying Ma. Hybrid index structures for location based web search. In Proceedings of the 14th ACM international conference on Information and knowledge management, Bremen, Germany, 2005. ACM Press.

S. Vaid, C. B. Jones, H. Joho, and M. Sanderson. Spatio-textual indexing for geographical search on the web. In Proceedings of 9th International Symposium on Spatial and Temporal Databases, 2005.

Bruno Martins, MSrio J. Silva, and Leonardo Andrade. Indexing and ranking in geo-ir systems.In Proceedings of the 2005 workshop on Geographic information retrieval, pages 31-34, Bremen, Germany, 2005.