Summary of Efficient Retrieval of the Top-k Most Relevant Spatial Web Objects

By S. Xiao Fernández Marín

1 Summary

This paper describes how to get the top-k most relevant documents or objects taking into account the spacial (geo-references points) and text relevancy. This is achieved using the *location-aware top-k text retrieval* (LkT queries).

They have based the text search in inverted indexes, consisting in mapping from words. Very used in full search like for example Google. There are four types of ways for looking for the top.k most relevant documents:
- The IR-trees, an R-tree but with inverted index.
- The DIR-trees, for also taking into account the document similarity.
- CIR-trees, for adding more information, having a more specific search.
- CDIR-trees, a combination of the last two trees.

IR-Trees: It is the most used index for spacial queries, used as R-trees but with inverted indexes. In this case, each leaf node contains a pointer to an inverted file containing vocabulary for all terms of a collection and a set of posting lists. The non-leaf nodes have entries as *(cp, rectangle, cp.di)*. *cp* as the address of the child, *rectangle* as the minimum bounding rectangle and *cp.di*, the identifies of a pseudo document.

DIR-Trees: Enhances the documents similarities, not only the geometry.

CIR-Trees: The data is divided into clusters and from this, a file is formed with the information of that cluster. The entries are in the way of *(O, O.rectangle, O.doc, O.C)*, where O is the cluster object, and $O.C$ is a label indicating to which cluster O belongs to and $O.doc$ is for the inverted file.

The experimental study shows that when varying k the runtime is proportional to the n^o of page accesses. Varying the number of keywords, the IR-tree and the DIR-tree improve the query performance. Varying α the IR-tree are better for large α and DIR-tree performs better for small α.

Varying the data size, runtime and I/O increases with the size and the indexes are more or less linearly and varying β, DIR-trees perform better when big β.

2 Questions not answered by this text

The authors leave some ideas that this paper can help with, like for example ”develop algorithms for other type of queries” or understanding ”how the top-k queries considered can best be processed if the spatial objects are constrained to a road network.”

3 What has changed since this was written

This algorithm has also been used for the collective spatial keyword querying for pruning the search space problem.