Text Search with Spatial Constraints

Amber Han and Bradford G. Nickerson

Faculty of Computer Science, University of New Brunswick, Fredericton, New Brunswick, Canada

Motivation

Improve search engines to better support I/O-efficient **text + spatial** queries.

Sample Data

Records consist of plain text and associated spatial information in a database

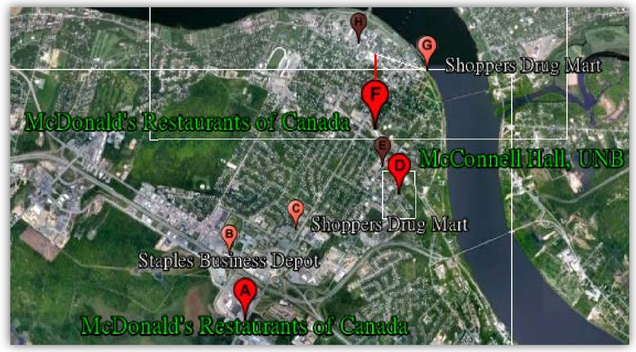
Record 1	Record 4
McDonald's Restaurants of Canada	Shoppers Drug Mart
(45.934516N, 66.663308W) **A**	(45.942639N, 66.655147W) **C**
(45.961817N, 66.643622W) **F**	(45.961817N, 66.643622W) **G**
	(45.976414N, 66.649003W) **I**
Record 2	**Record 5**
McConnell Hall, University of New Brunswick	Staple Business Depot
(45.946419N, 66.639297W) **D**	(45.939778N, 66.662633W) **B**
Record 3	**Record 6**
Head Hall, University of New Brunswick	Victory Meat Market Ltd
(45.949961N, 66.641711W) **E**	(45.962739N, 66.645572W) **H**

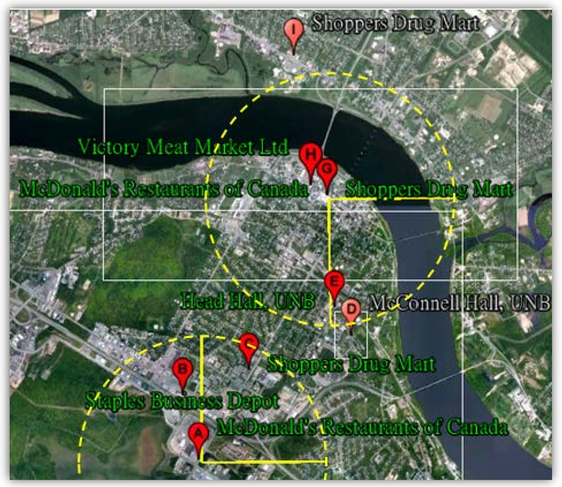
Three query types

Q1 = (t): search returns a ranked list of items matching the search string t, along with their associated spatial information (e.g. latitude, longitude).

- Example query: Q = ("Mc")

Records A, F, D are returned.





Q2 = (t, r): search returns a ranked list of documents with at least one spatial component having its location falling within the circle of radius r centred at the position p of the ranked documents matching t.

- Example query: Q2 = ("McDonald", 1.5km)

Records A, F, G, H, B, C, E are returned.

Q3 = (p, r): search returns a ranked list of documents with at least one spatial component having its location falling within the circle of radius r centred at position p.

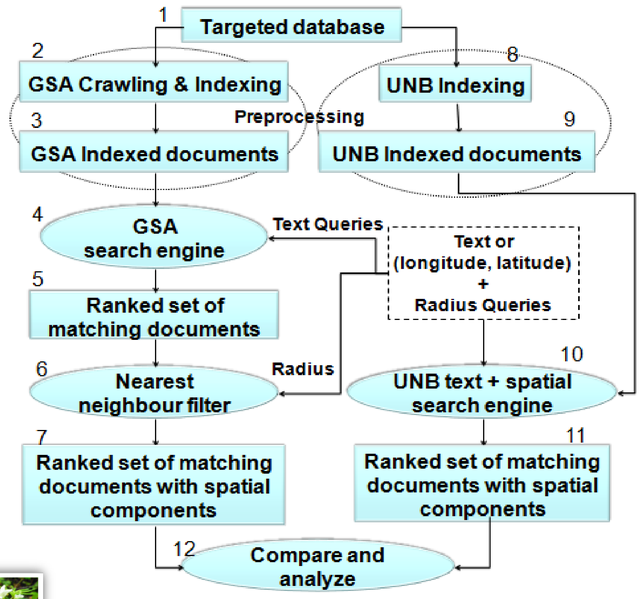
- Example query: Q3 = (position P = (45.952567N, 66.646001W), 1.5km)

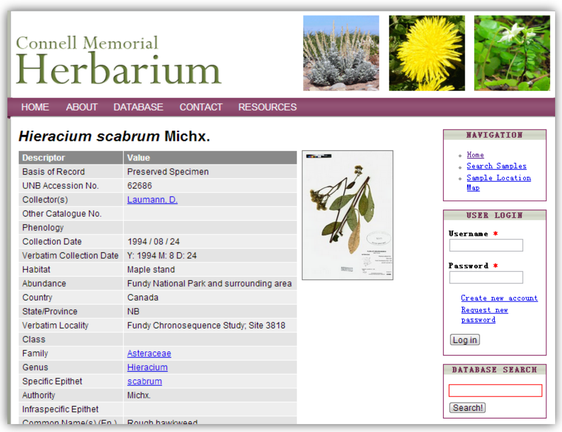
Records E, D, F, G, H, C are returned.

In Q2 and Q3 search results, the points nearer to the point(s) of interest are ranked higher

Test Methodology

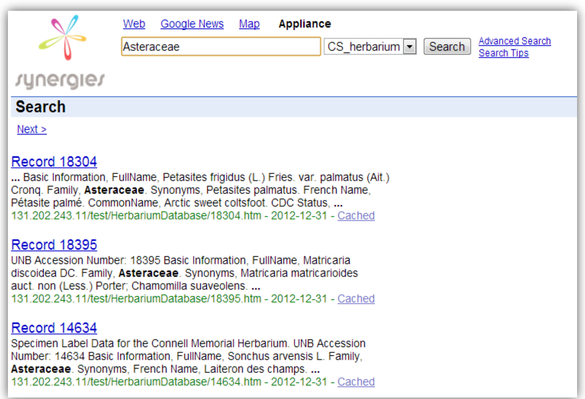
The system will be tested on the UNB Connell Memorial Herbarium database (N = 40,791). The performance of the system will be compared with the Google Search Appliance (GSA) on the targeted data set.



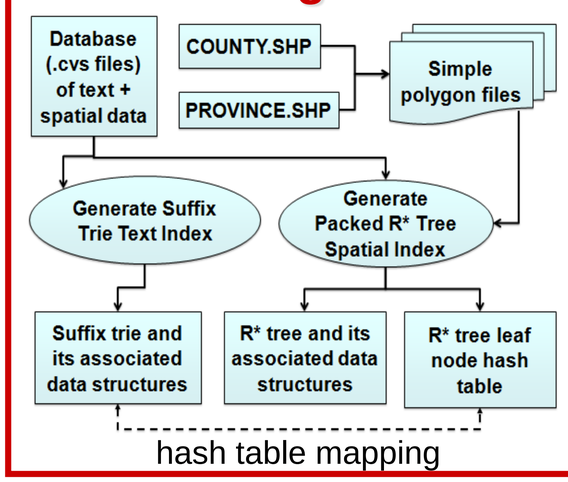


Example record from the UNB Connell Memorial Herbarium database with the query string "Asteraceae"

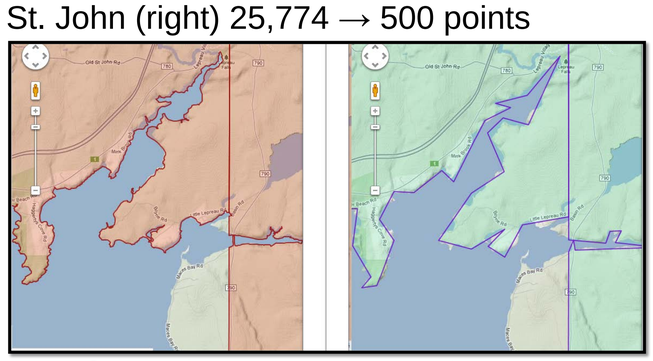
Example result from the GSA on the UNB Herbarium data set with query string "Asteraceae"



The indexing scheme



To reduce space, we simplify the polygons using the **PackPolygon** algorithm.
Charlotte (left) 27,756 → 500 points
St. John (right) 25,774 → 500 points



Original Packed

Sponsored by:

UNB Faculty of COMPUTER SCIENCE

NSERC CRSNG