Cloud based implementation: Geo-Crawler

Prepared by

Deepak Punjabi (*15IT60R17*)
Bhumi Faldu (*15IT60R18*)
Mayank Gautam (*15IT60D04*)

Under Guidance of

Professor Soumya K. Ghosh

Building a cloud based implementation for a spatial web crawler that crawls through the web to find and store web feature services classify and index them for efficient retrieval.

“

Problem Definition

Objectives

- Implementing a **cloud based architecture** to build and efficient web crawler.
- **Building a spatial web crawler** using *WFS* based on *OGC* standard.
- Building a ***domain ontology*** with spatial *feature type*.
- **Semantic matching** using *ontology* and indexing of geo-servers with offered *feature type* reference.
- Performing experiment with test seed *URLs* and **analysing the performance** of the crawler in terms of accurate semantic annotations.

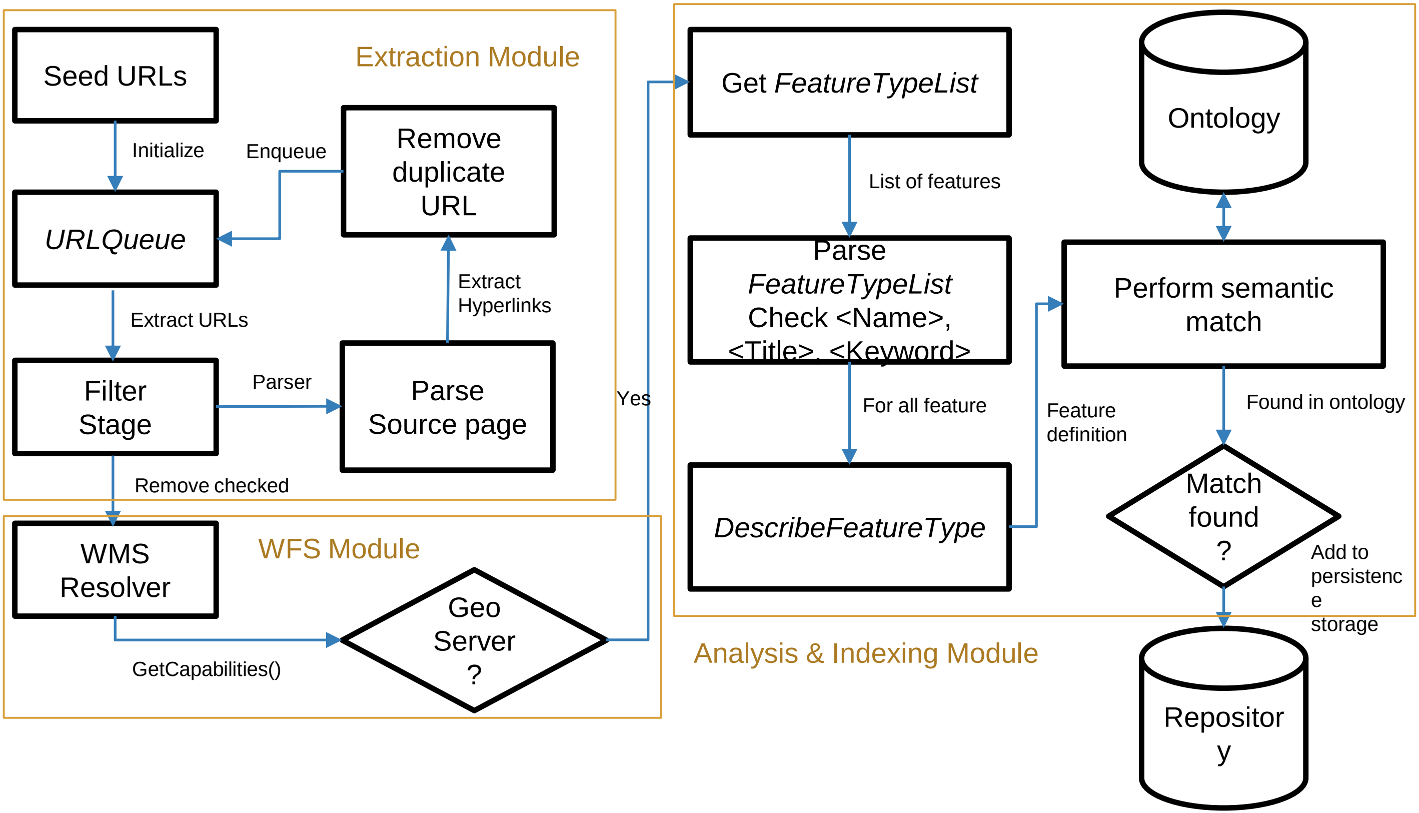
Solution Methodology

Master-Worker Approch

- ❑ One master, multiple worker
- ❑ Master manages workers
- ❑ Task is divided into multiple smaller tasks by master
- ❑ Master assigns task to worker and keep track of the work state
- ❑ All workers run this task parallel
- ❑ Master does the synchronization between the workers

Map-Reduce Approach

- ❑ Framework includes two types of nodes, mapper nodes and reducer nodes
- ❑ Mapper node crawls through URLs to fetch and filter URLs and store them into list
- ❑ Reducer node checks if the URL belong to a geo-server
- ❑ If yes, then it extracts the feature metadata from the server and stores it into the repository



Spatial Crawler Architecture

Result Metric

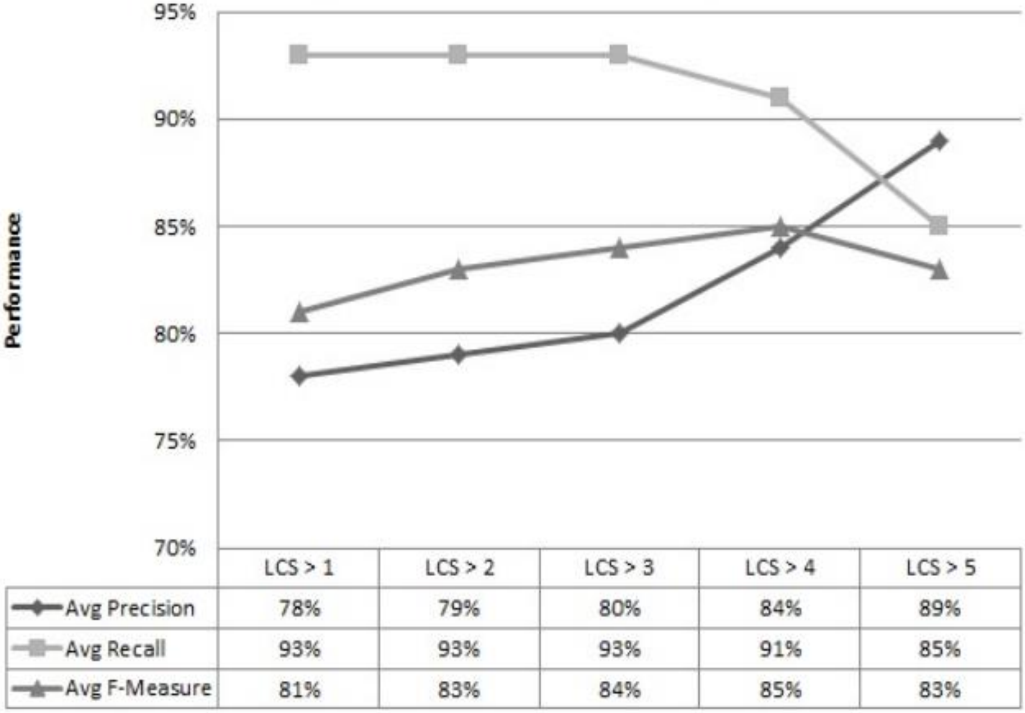
- $precision = \frac{(Number\_of\_relevant\_geoservers\_found)}{(Total\_Number\_of\_geoservers\_found)} * 100\%$

- $recall = \frac{Number\_of\_relevant\_geoservers\_found\_in\_search}{Total\_Number\_of\_existing\_relevant\_geoservers} * 100\%$

- $F1 = 2 * \frac{(precision * recall)}{(precision + recall)}$

Final score is calculated by taking average over all *feature types*.

Performance based on LCS threshold



	LCS > 1	LCS > 2	LCS > 3	LCS > 4	LCS > 5
◆ Avg Precision	78%	79%	80%	84%	89%
■ Avg Recall	93%	93%	93%	91%	85%
▲ Avg F-Measure	81%	83%	84%	85%	83%

source: patil et al, springer 2014

Conclusive Thoughts

- In the recent era, need of storing and retrieving sptial data and feature is a necessity
- To retrieve sptial data, it is necessary to understand it's feature and operations
- To cater the need of spatial data, crawler based approach is implemented
- Cloud based approch can be followed for the implementation using both master slave and map reduce approach.
- Massive parallelization can be applied to cater the peformance need for the crawler.
- Cloud based approach provides cost effective and scalable architecture. It satisfies the economy of scale.

References

I. Patil, Sonal, Shrutilipi Bhattacharjee, and Soumya K. Ghosh. "**A spatial web crawler for discovering geo-servers and semantic referencing with spatial features.**" Distributed Computing and Internet Technology. Springer International Publishing, 2014. 68-78.

II. Li, Wenwen, Chaowei Yang, and Chongjun Yang. "**An active crawler for discovering geospatial web services and their distribution pattern–a case study of OGC web map service.**" International Journal of Geographical Information Science 24.8 (2010): 1127-1147.

III. Jiang, Jun, Chong-jun Yang, and Ying-chao Ren. "**A spatial information crawler for opengis wfs.**" Sixth International Conference on Advanced Optical Materials and Devices. International Society for Optics and Photonics, 2008.

IV. Marc Najork. "**Web crawler architecture.**" Microsoft Research.

V. Ahlers, Dirk, and Susanne Boll. "**Location-based Web search**." The Geospatial Web. Springer London, 2009. 55-66.

VI. Li, W., et al. "**Semantic-based web service discovery and chaining for building an Arctic spatial data infrastructure.**" Computers & Geosciences 37.11 (2011): 1752-1762.

VII. Suakanto, Sinung, et al. "**Building crawler engine on cloud computing infrastructure.**" Cloud computing and social networking (ICCCSN), 2012 international conference on. IEEE, 2012.

VIII. Bahrami, Mehdi, Mukesh Singhal, and Zixuan Zhuang. "**A cloud-based web crawler architecture.**" Intelligence in Next Generation Networks (ICIN), 2015 18th International Conference on. IEEE, 2015.

Thanks!

Any questions ?