Mining and Utilizing Dataset Relevancy from Oceanographic Dataset (MUDROD) Metadata, Usage Metrics, and User Feedback to Improve Data Discovery and Access

Chaowei (Phil) Yang, Yongyao Jiang, Yun Li,
George Mason University

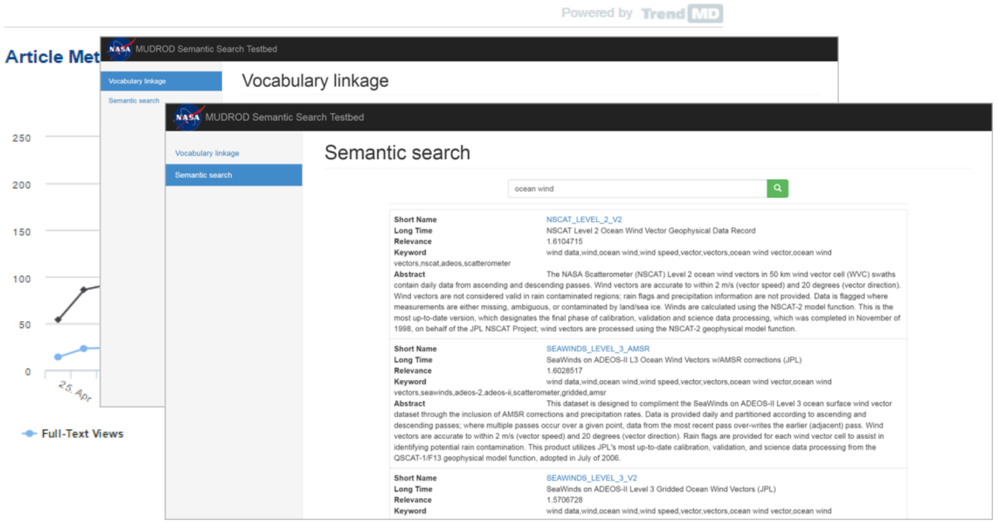
Edward M Armstrong, Thomas Huang, David Moroni,
Chris Finch, Lewis Mcgibbney, JPL, NASA

Mining and Utilizing Dataset Relevancy from Oceanographic Data (MUDROD)

PI: Chaowei (Phil) Yang, George Mason University

Objective

- Improve data discovery, selection and access to NASA Observational Data.
    - Intuitive interface to federated data holdings.
    - Enable new user communities to discover and access data for their projects.
    - Reduce time for scientists to discover, download and reformat data.
    - Implement extensible ontology framework.
    - Improve discovery accuracy of oceanographic data
    - Foundation for Managing Big Data.
- Demonstrate MUDROD at PO.DAAC.



Approach:

- Setup collaboration, testing environment.
- Design MUDROD knowledge base system.
- Develop P.O.DAAC user service.
- Update semantic search and conduct alpha testing.
- Integrate MUDROD alpha into P.O.DAAC.
- Enhance knowledge base, to include GCMD.
- Integrate selected datasets from ECHO.
- Outreach to GEO
- Demonstrate Prototype.

Key Milestones

• Start	06/15
• Identify Use cases	01/16
• Design search, query, reasoning engine	03/16
• Ontological System Implementation	07/16
• Complete Beta test at P.O. DAAC	12/16
• Integrated test	02/17
• PO.DAAC metadata discovery Demo (TRL 7)	05/17

CoIs: T. Huang, D. Moroni, E. Armstrong, JPL;

TRL_{in} = 5, $TRL_{current}$ = 6

Agenda

- Context
- Objectives
- Data
- Mining concept linkage multi-sources
- Integration
- Demo

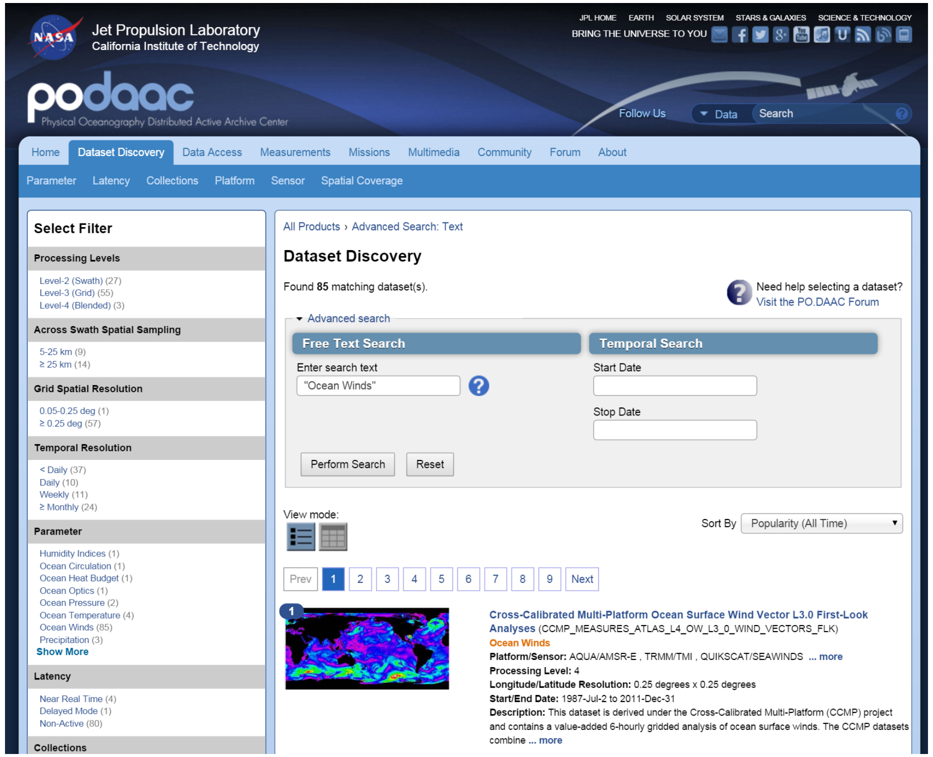
Data Discovery Problems

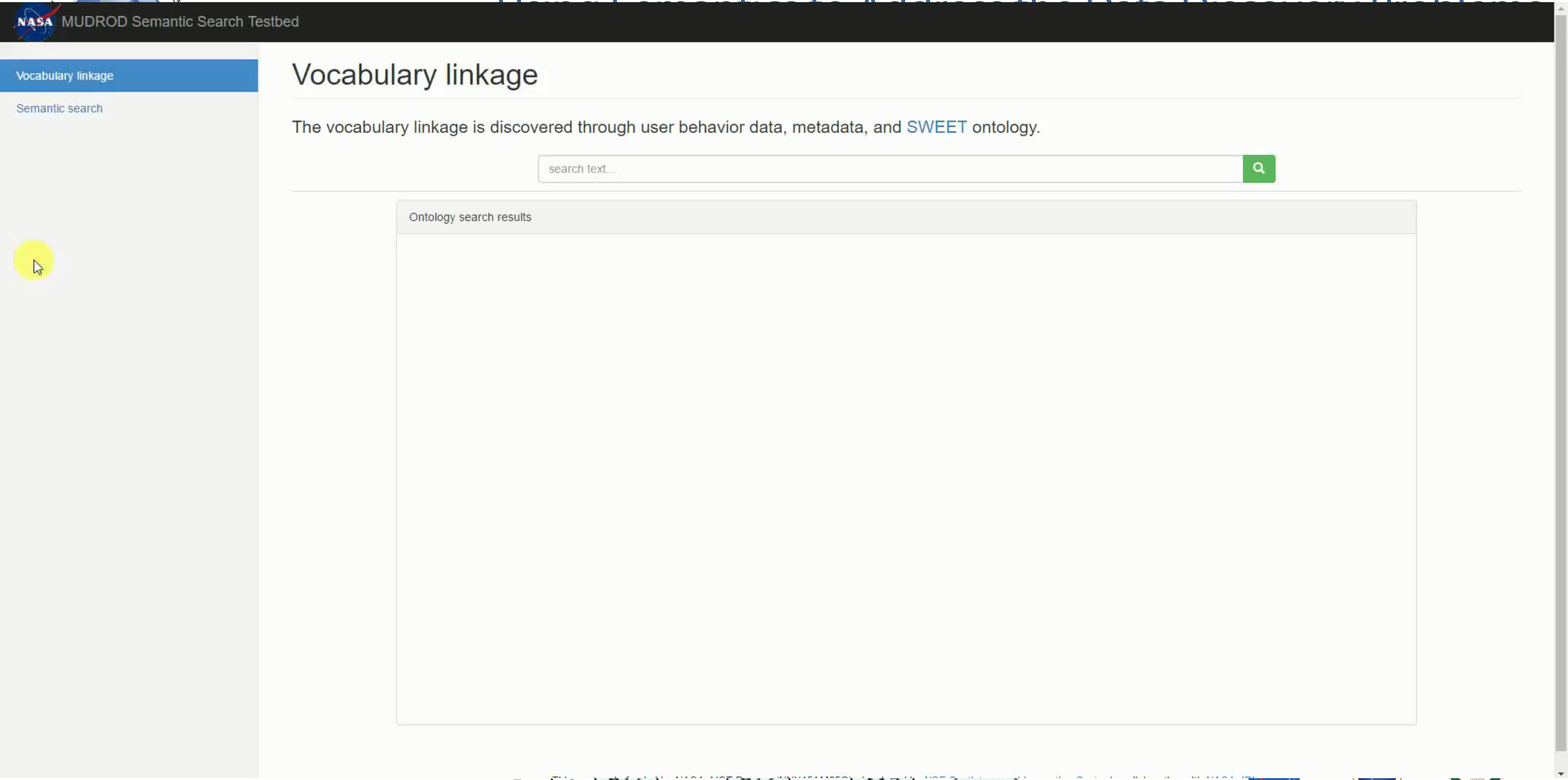
- Keyword-based matching (traditional search engines)
  - User query: **ocean wind**
  - Final query: ocean AND wind

- Reveal the real intent of user query
  - ocean wind = "ocean wind" OR "greco" OR "surface wind" OR "mackerel breeze" …

- UWG Recommendation 2014-07

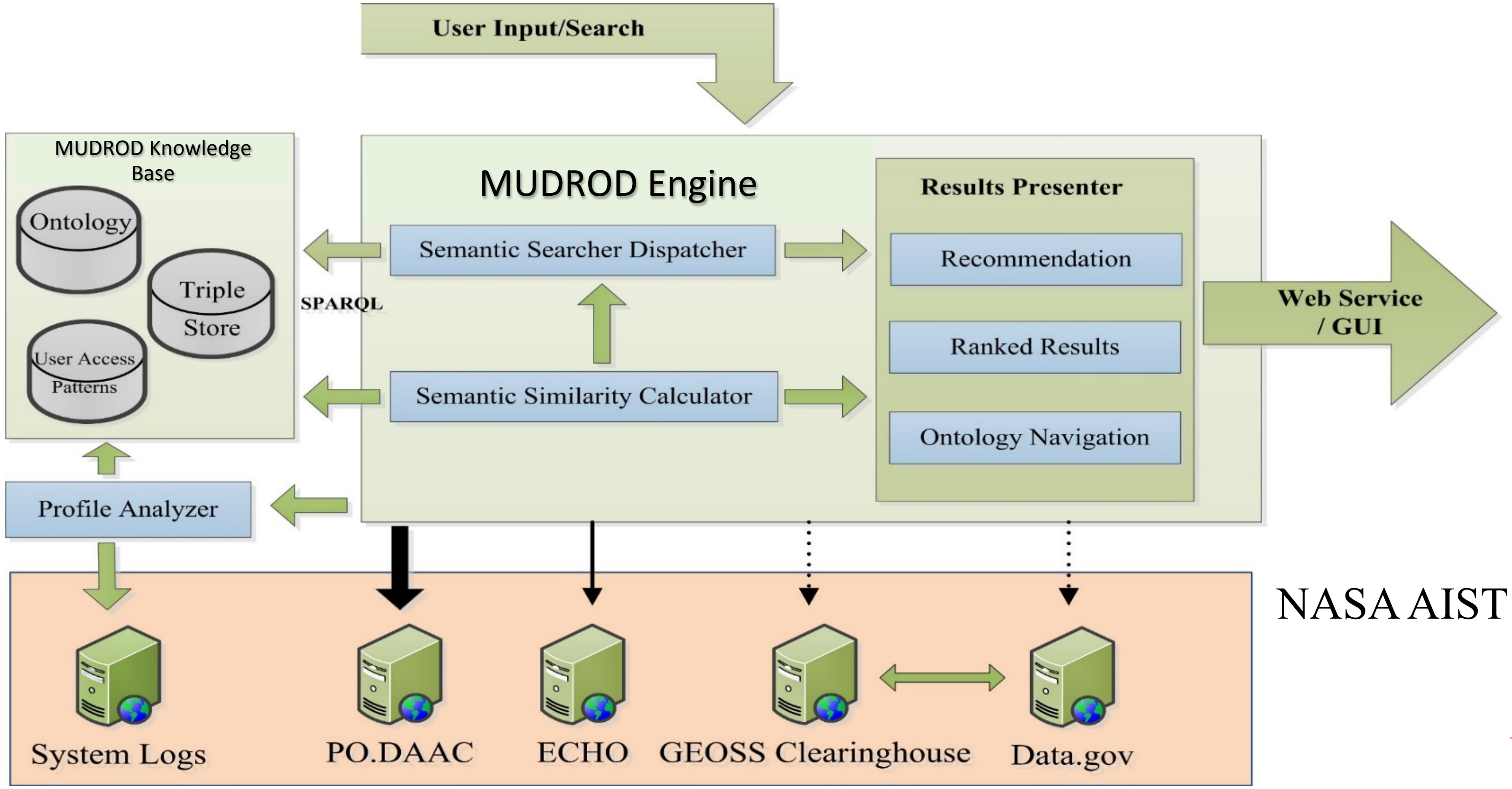
- ESDSWG Search Relevance




MUDROD Semantic Search Testbed
Vocabulary linkage
Semantic search
Vocabulary linkage
The vocabulary linkage is discovered through user behavior data, metadata, and SWEET ontology.
search text...
Ontology search results

Objectives

- Analyze web logs to discover user knowledge of the connections between datasets and keywords
- Construct knowledge base by combining semantics and profile analyzer
- Improve data discovery with 1) better ranked results; 2) recommendation; and 3) Ontology navigation



NASA AIST (NNX15AM85G)

Research workflow

Mining PO.DAAC metadata

Extract SWEET ocean concepts

Session reconstruction from logs

Mining user query history

Mining user clicking behavior

Integrate mining results into MUDROD knowledge base

Build a sustainable MUDROD knowledge base (ontology, triple store, pattern, etc.)

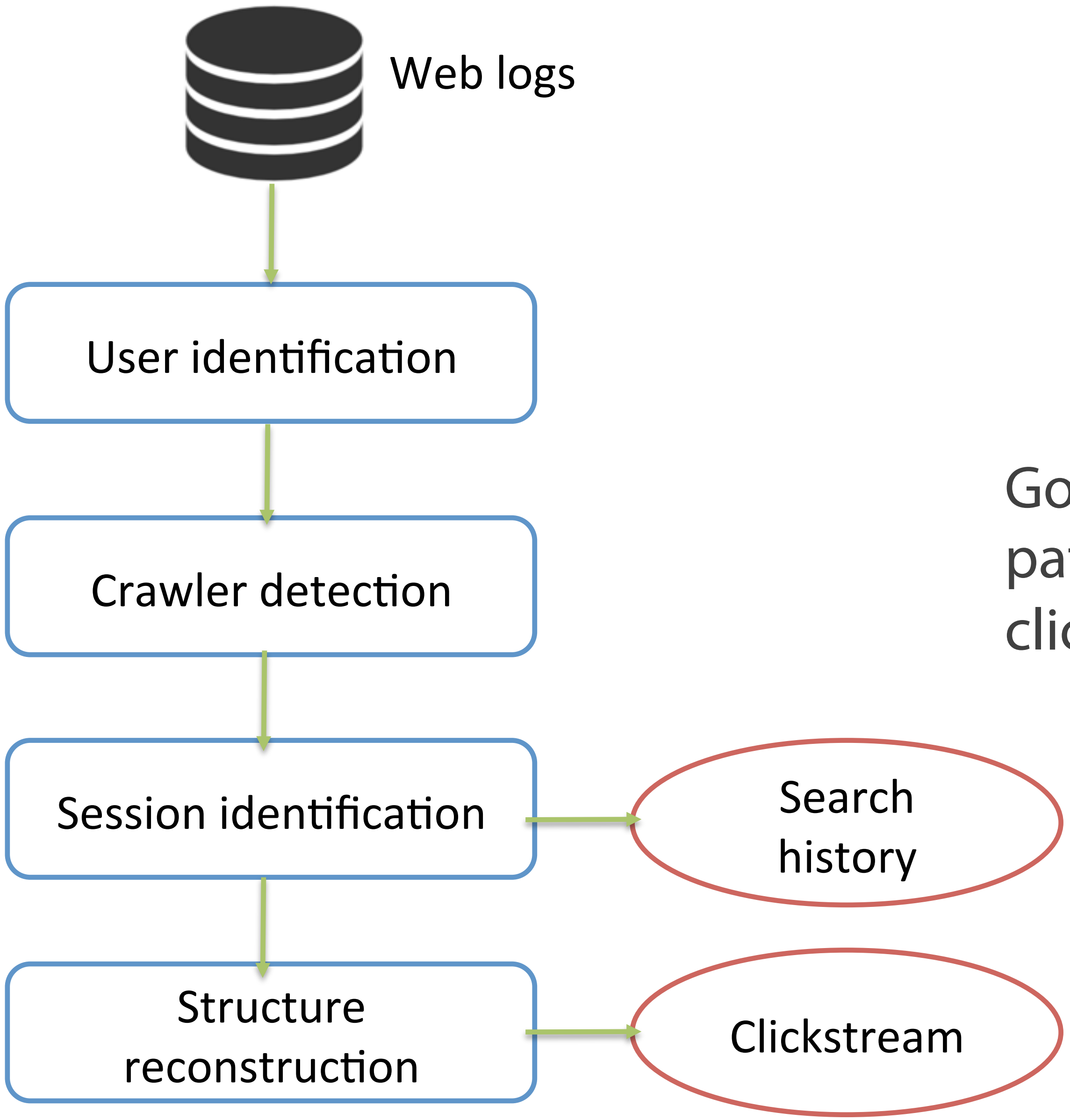
3: Data

- Log files of 2014 from PO.DAAC
- Access log (HTTP), FTP
- Requests sent from client, recorded by server

68.180.228.99 - - [31/Jan/2015:23:59:13 -0800] "GET /datasetlist/... HTTP/1.1" 200 84779 "/ghrsst/" "Mozilla/5.0 ..."

Client IP: 68.180.228.99
Request date/time: [31/Jan/2015:23:59:13 -0800]
Request: " GET /datasetlist/... HTTP/1.1 "
HTTP Code: 200
Bytes returned: 84779
Referrer/previous page: “/ghrsst/"
User agent/browser: "Mozilla/5.0 ...

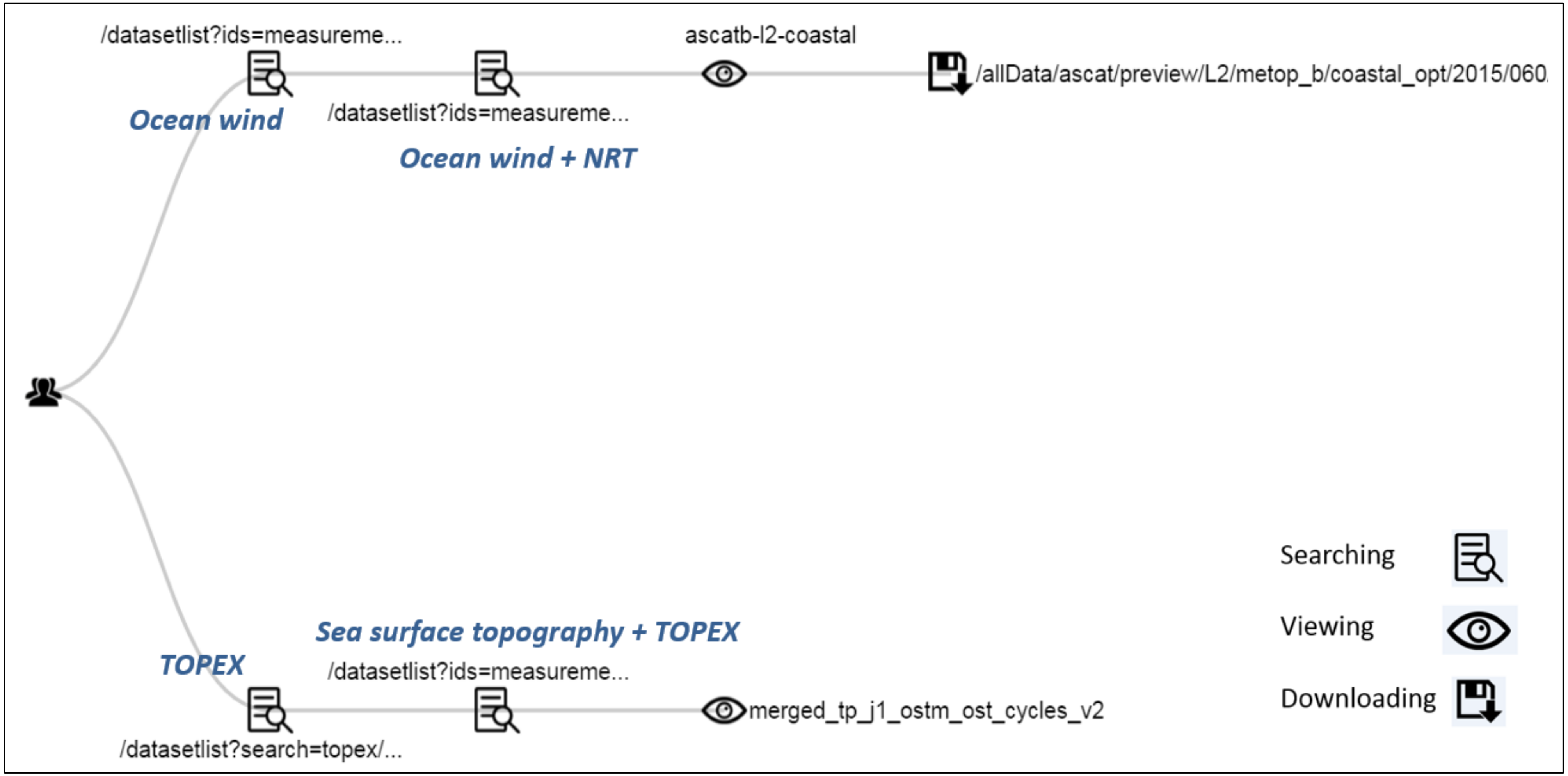
Data preparation



Goal: reconstruct user browsing pattern (search history & clickstream) from a set of logs

Additional steps include: word normalization, stop words removal, and stemming (NLP techniques)

Reconstructed session structure



Data preparation results

1. User search history

```
{
    "User A": [
        "modis",
        "sst",
        "ocean winds",
        "surface wind"
        ...
    ]
}
```

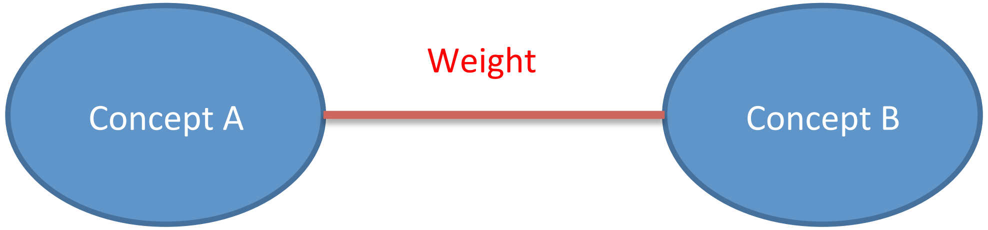
2. Click stream/through

```
{
    "Query": "sst AND Json-1",
    "View": "navo-l2p-avhrr19_g",
    "Download": "navo-l2p-avhrr19_g"
}
```

Technical approach of discovering linkage

Processed data	Hypothesis/Rationale	Approach
User search history (past searched queries)	The set queries conducted by each individual user are statistically related.	Binary cosine similarity (query- user)
Clickstream (linkage between query, viewing, downloading)	Similar query can result in statistically similar clicking behavior	TF-IDF normalization, SVD, cosine similarity (query – data)
Metadata (DIF, extracted properties)	Similar terms are more likely to be in the same metadata	Latent semantic indexing (term- data)
Existing ontology (SWEET)	"SubClassOf" (Hyponymy) and "equivalentClass" (Synonym)	Path-based similarity (directed)

- All four results could be converted to



- Problem:
  - Neither search history nor clicking behavior are perfect, due to the processing uncertainty in data, hypothesis and method
  - Metadata and existing ontology might have unknown terms to search engine end users
  - Better determine the final similarity

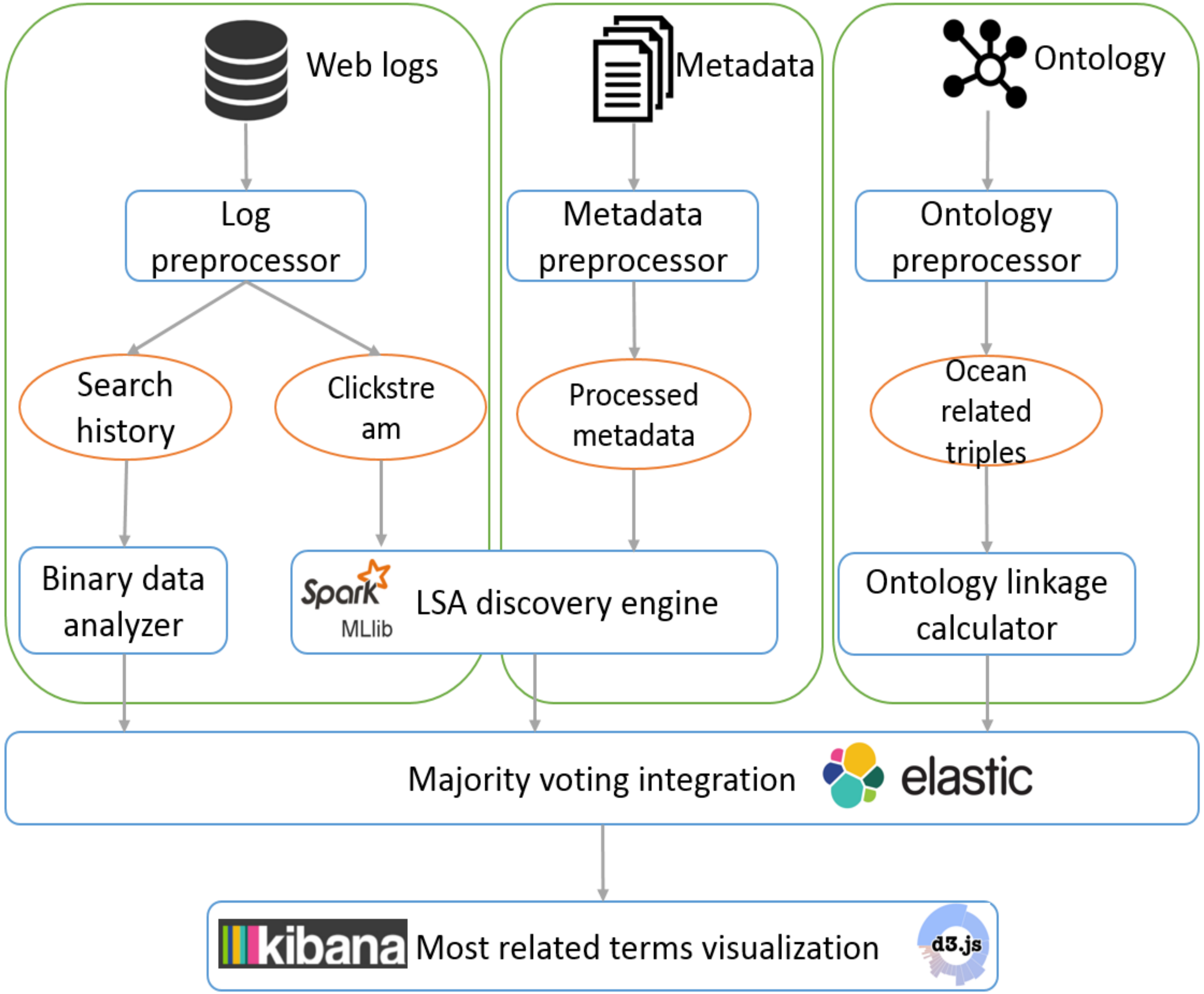
Solution: majority voting rule

$$sim(X,Y) = \frac{\sum_i sim_i \cdot w_i}{\sum_i w_i} + (\sum_i w_i - \theta) \cdot \beta \qquad (11)$$

Where method i is the method that has the linkage of (X,Y), w_i is the weight of method i, sim_i is the similarity of (X,Y) in method i, θ is the threshold that represents the minimum sum of methods weights that makes the linkage a majority, and β is a constant that represents the majority rule change rate.

- Weighted average similarity of different models
- Strong relationships agreed upon by more sides would become stronger

System architecture

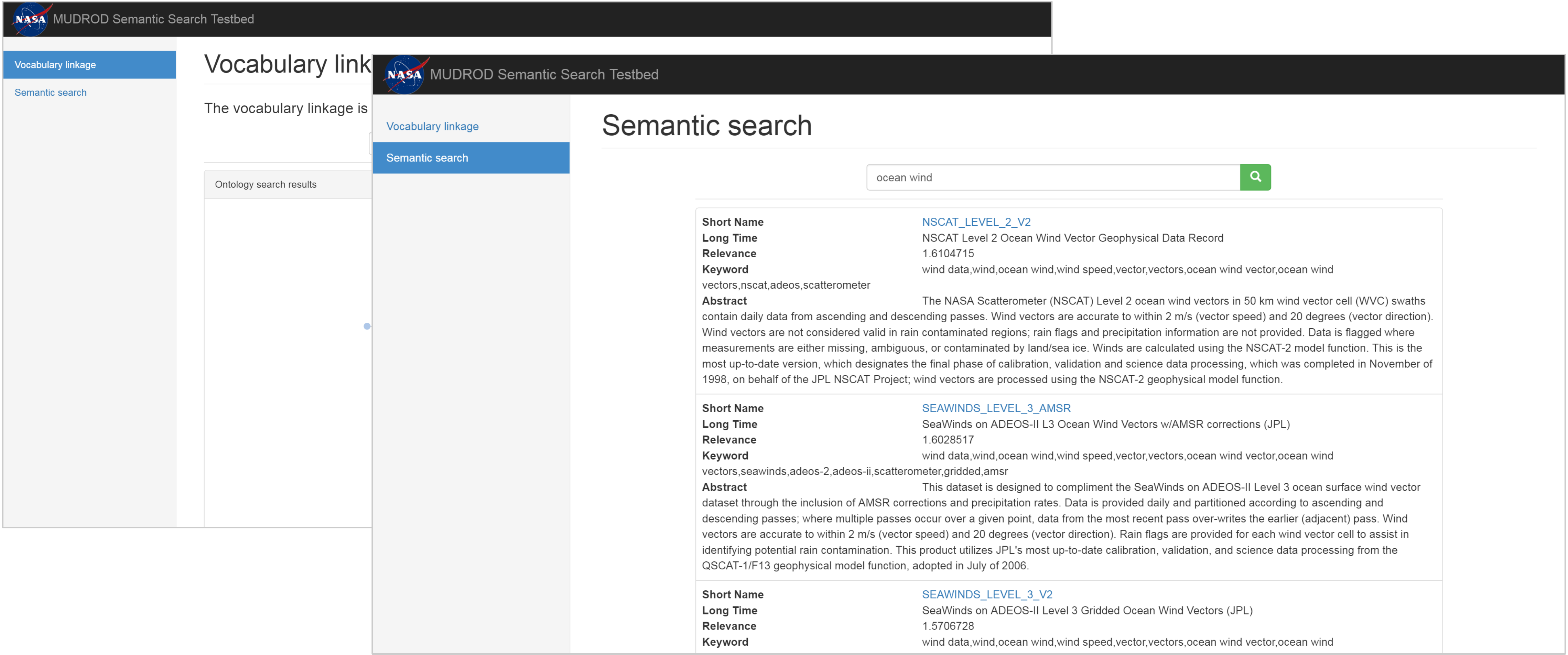


Results and evaluation

	Search history	Clickstream	Metadata	SWEET	Integrated list
ocean temperature	sea surface temperature(0.66), sea surface topography(0.56), ocean wind(0.56), aqua(0.49), ocean circulation(0.49)	sea surface temperature(0.94), sst(0.94), group high resolution sea surface temperature dataset(0.89), ghrsst(0.87), caspian sea(0.74)	sst(0.96), ghrsst(0.77), sea surface temperature(0.72), surface temperature(0.63), reynolds(0.58)	None	sst(1.0), sea surface temperature(1.0), ghrsst(1.0), group high resolution sea surface temperature dataset(0.99), reynolds sea surface temperature(0.74)

Sample group	Overall accuracy
Most popular 10 queries	88%
Least popular 10 queries	61%
Randomly selected 10 queries	81%

Demo



http://52.70.209.189:8080/ontology/search.html

Cloud computing experiment

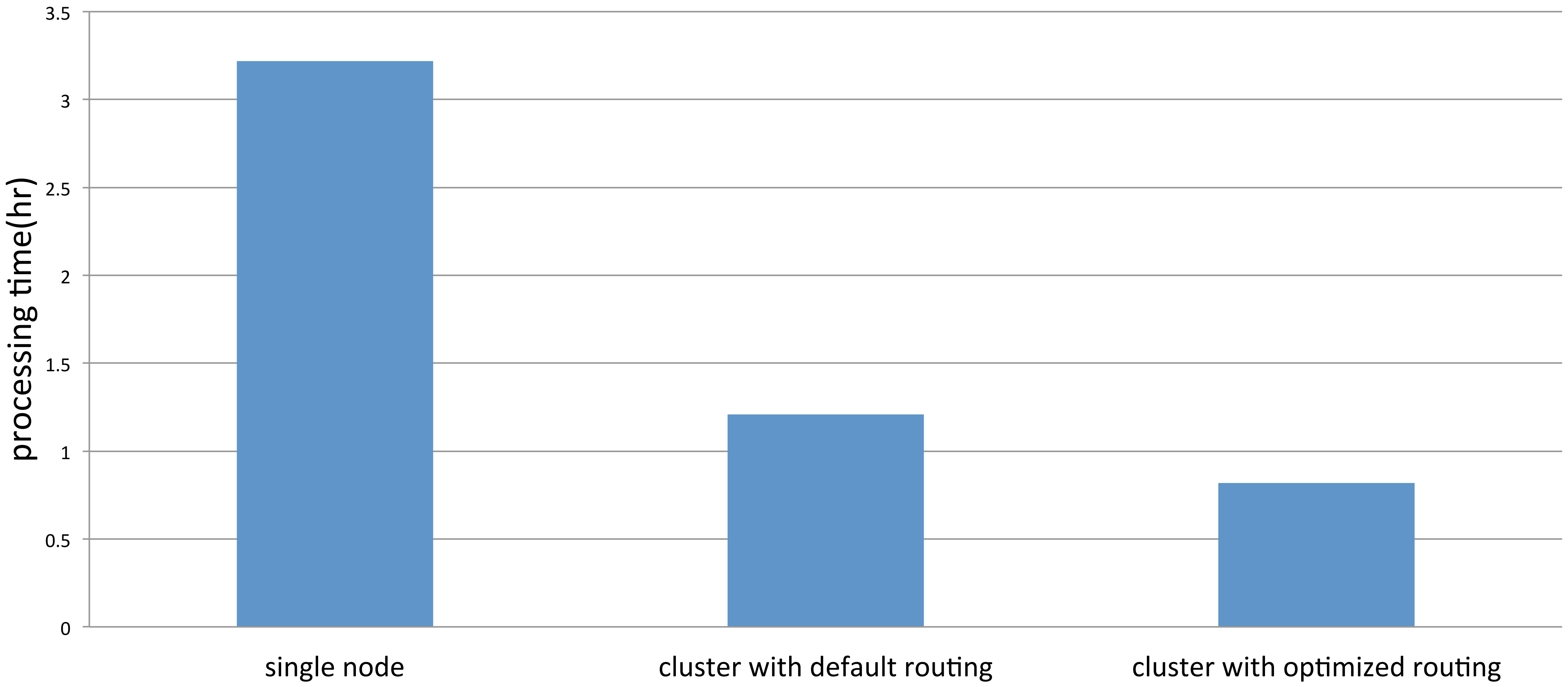
Data: 32.8GB, 154,969,689 (0.15 billion) lines in total (PODAAC logs of 2014)
VMs: 4 CPU cores, 16G memory, Clock Speed 2.4GHz

Experiment	VMs	Cluster node	Cluster mode	Log mining client	Time
1	1	1	Default routing	1	3 hours and 13 minutes
2	4	4	Default routing	4	1 hour and 12 minutes
3	4	4	Optimized routing	4	49 minutes

- Default routing gives an even distribution of documents across the entire cluster using document ID.
- Custom routing assigns documents to a node using value of IP filed, in order to reduce network transportation cost.

Experiment results

Processing time of 3 experiments



Processed data: 32.8GB, 154,969,689 (0.15 billion) lines in total

1st Year Phased Progress

Component testing, deployment, PO.DAAC data, PO.DAAC UWG/Scientists

1. Quarter 1: Setup the collaboration and testing environment,
2. Quarter 2: Design MUDROD knowledge base, engine and GUI
3. Quarter 3: Develop PO.DAAC user search and download profile service
4. Quarter 4: Update the semantic search based on the MUDROD system design and conduct alpha development testing.

2nd Year Phased Progress

CMR and other DAACs, PO.DAAC UWG/scientists, Open Source consideration

1. Quarter 1: Parallelizing mining using cloud and improving relevance ranking; Integrate MUDROD alpha into PO.DAAC Labs and demonstration to PO.DAAC UWG
2. Quarter 2: MUDROD beta testing and enhance knowledge base to include Non-NASA ocean taxonomies, hosted by the NASA CMR
3. Quarter 3: Integrate selected CMR metadata and CMR user profile statistics for CMR integration and conduct operation testing
4. Quarter 4: demonstrate the developed prototype system at PO.DAAC, CMR, and CLH operations.

Next Quarter/6 Month plan

– Improve ranking based on the vocabulary linkage and user behavior
– Build MUDROD ontology
– Ontology navigation and recommendation
– ESIP Testbed Evaluation
– Integrate alpha into PO.DAAC Labs

– **Open source**
– **Test and apply as cross-infrastructure capability (e.g., supporting solr)**

Test plan with ESIP

- Put projects into the ESIP Testbed
- Prepare installation and user guide
- Contact with Annie Burgess(2-3 times in the following 3 weeks)
  - how evaluators will access MUDROD
  - cyberinfrastructure required for evaluators to access MUDROD
  - current project TRL
  - discuss evaluation objectives and process
- Select evaluators (by ESIP)
- Create test plans (by evaluators and PI)
- Conduct an independent evaluation of the TRL and usability(by evaluators)
  - milestone completion review
  - TRL objective completion review
  - Assess MUDROD using the TRL Evaluation Structure
- Submit final report to ESIP (by evaluators)
- ESIP confirm with PI, then submit to Mike Little

- https://docs.google.com/document/d/14dp3Vkjp5A0vRpcOrWfzxLn3BEzl9Ks7VBw7sUwCsc4/edit?usp=sharing

Component	Current TRL	Target TRL	Description
Semantic search engine			
Semantic Search Dispatcher	6	7	
Semantic Similarity Calculator	6	7	
Recommendation	N/A	7	
Ranked Results	6	7	
Ontology Navigation	6	7	
Knowledge base			
Ontology	N/A	7	
Triple Store	N/A	7	
User Access Pattern	6	7	
Vocabulary linkage discovery engine			
Profile analyzer	6	7	
Metadata analyzer	6	7	
Ontology linkage calculator	6	7	
GUI			
Vocabulary linkage visualization tool	6	7	
Semantic search presenter	6	7	

Publications and Presentations

- **Papers**
- Jiang, Y., Y. Li, C. Yang, E. M. Armstrong, T. Huang & D. Moroni (2016) Reconstructing Sessions from Data Discovery and Access Logs to Build a Semantic Knowledge Base for Improving Data Discovery. *ISPRS International Journal of Geo-Information*, 5**,** 54. http://www.mdpi.com/2220-9964/5/5/54#stats
- Jiang, Y., Y. Li, C. Yang, K. Liu, E. M. Armstrong, T. Huang & D. Moroni (2016) A Comprehensive Approach to Determining the Linkage Weights among Geospatial Vocabularies - An Example with Oceanographic Data Discovery. (drafted in review)
- Y. Li, Jiang, Y., C. Yang, K. Liu, E. M. Armstrong, T. Huang & D. Moroni (2016) Leverage cloud computing to improve data access log mining. (in progress)

- **Conference presentations**
- Yang C., Jiang Y., L Y., Armstrong E., Huang T., and Moroni D., 2015. “Utilizing Advanced IT Technologies to Support MUDROD to Advance Data Discovery and Access”, AGU, San Francisco, CA.
- Yang C., Jiang Y., L Y., Armstrong E., Huang T., and Moroni D., 2016. “Mining and Utilizing Dataset Relevancy from Oceanographic Dataset (MUDROD) Metadata, Usage Metrics, and User Feedback to Improve Data Discovery and Access”, ESIP winter meeting 2016, Washington D.C.
- Jiang Y., Yang C., L Y., Armstrong E., Huang T., and Moroni D., 2016. “A Comprehensive Approach to Determining the Linkage Weights among Geospatial Vocabularies - An Example with Oceanographic Data Discovery”, AAG 2016, San Francisco, CA.
- Yang C., Jiang Y., L Y., Armstrong E., Huang T., and Moroni D., 2016. “Mining and Utilizing Dataset Relevancy from Oceanographic Dataset (MUDROD) Metadata, Usage Metrics, and User Feedback to Improve Data Discovery and Access”, PO.DAAC UWG, Pasadena, CA.

Acknowledgements

1. NASA AIST Program (NNX15AM85G)
2. PO.DAAC SWEET Ontology Team (Initially funded by ESTO)
3. Hydrology DAAC Rahul Ramachandran (providing the earlier version of NOESIS)
4. ESDIS for providing testing logs of CMR
5. All team members at JPL and GMU

References

- Abrol, M. & B. Johnson. 2001. Adaptive document ranking method based on user behavior. Google Patents.
- Agichtein, E., E. Brill & S. Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 19-26. ACM.
- AlJadda, K., M. Korayem, T. Grainger & C. Russell. 2014. Crowdsourced query augmentation through semantic discovery of domain-specific jargon. In *Big Data (Big Data), 2014 IEEE International Conference on*, 808-815. IEEE.
- Arguello, J. (2013) Vector Space Model. *Information Retrieval September*, 25.
- Berners-Lee, T., J. Hendler & O. Lassila (2001) The semantic web. *Scientific american*, 284**,** 28-37.
- Blei, D. M., A. Y. Ng & M. I. Jordan (2003) Latent dirichlet allocation. *the Journal of machine Learning research*, 3**,** 993-1022.
- Casey, K. 2016. NOAA OneStop Data Discovery and Access Framework Project: Progress, Feedback, and Alignment with the USGEO Common Framework on Earth Observation Data. ESIP Summer Meeting 2016: ESIP.
- Christian, E. J. (2008) GEOSS architecture principles and the GEOSS clearinghouse. *Systems Journal, IEEE*, 2**,** 333-337.
- Díaz-Galiano, M. C., M. García-Cumbreras, M. T. Martín-Valdivia, A. Montejo-Ráez & L. Urena-López. 2007. Integrating mesh ontology to improve medical information retrieval. In *Advances in multilingual and multimodal information retrieval*, 601-606. Springer.
- De Lathauwer, L., B. De Moor & J. Vandewalle (2000) A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21**,** 1253-1278.
- Devarakonda, R., G. Palanisamy, J. M. Green & B. E. Wilson (2011) Data sharing and retrieval using OAI-PMH. *Earth Science Informatics*, 4**,** 1-5.
- Dumais, S. T. (2004) Latent semantic analysis. *Annual review of information science and technology*, 38**,** 188-230.
- Grossman, D. A. & O. Frieder. 2012. *Information retrieval: Algorithms and heuristics*. Springer Science & Business Media.
- Gui, Z., C. Yang, J. Xia, K. Liu, C. Xu, J. Li & P. Lostritto (2013) A performance, semantic and service quality-enhanced distributed search engine for improving geospatial resource discovery. *International Journal of Geographical Information Science*, 27**,** 1109-1132.
- Hua, X.-S., L. Yang, J. Wang, J. Wang, M. Ye, K. Wang, Y. Rui & J. Li. 2013. Clickage: Towards bridging semantic and intent gaps via mining click logs of search engines. In *Proceedings of the 21st ACM international conference on Multimedia*, 243-252. ACM.
- Jiang, Y., M. Sun & C. Yang (2016b) A Generic Framework for Using Multi-Dimensional Earth Observation Data in GIS. *Remote Sensing*, 8**,** 382.
- Jiang, Y., C. Yang, J. Xia & K. Liu. 2016c. Polar CI Portal: A Cloud-Based Polar Resource Discovery Engine. In *Cloud Computing in Ocean and Atmospheric Sciences*, ed. T. C. Vance, Merati, N., Yang, C., Yuan, M., 163-185. Academic Press.
- Jin, L., Y. Chen, T. Wang, P. Hui & A. V. Vasilakos (2013) Understanding user behavior in online social networks: A survey. *Communications Magazine, IEEE*, 51**,** 144-150.
- Konstan, J. A., B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon & J. Riedl (1997) GroupLens: applying collaborative filtering to Usenet news. *Communications of the ACM*, 40**,** 77-87.

Backups

User search history

- Hypothesis: the more frequent two queries co-occur in distinct users' search history, the more similar they are.

- Filter out rarely searched queries.
- Create query – user matrix
- Calculate binary cosine similarity

$$sim\,(t,s) = \frac{|t \cap s|}{\sqrt{|t| \cdot |s|}}$$

	User 1	User 2	…	User n
Query 1	1	1	…	1
Query 2	1	1	…	1
Query 3	0	0	…	0
…	…	…	…	…
Query m	0	0	…	1

Similar

Not

Clickstream

- Hypothesis: if two queries are similar, the data that get viewed and downloaded would be similar

- Filter out rarely searched queries.
- Create query – data matrix
- TF-IDF normalization, SVD
- Calculate cosine similarity

$$sim(t,s) = \frac{\vec{t} \cdot \vec{s}}{|\vec{t}| * |\vec{s}|}$$

Similar

Not

	Data 1	Data 2	…	Data n
Query 1	5	5	…	1
Query 2	10	10	…	0
Query 3	0	1	…	9
…	…	…	…	…
Query m	3	2	…	2

Metadata

- Hypothesis: semantically related terms tend to appear in the same document more frequently.

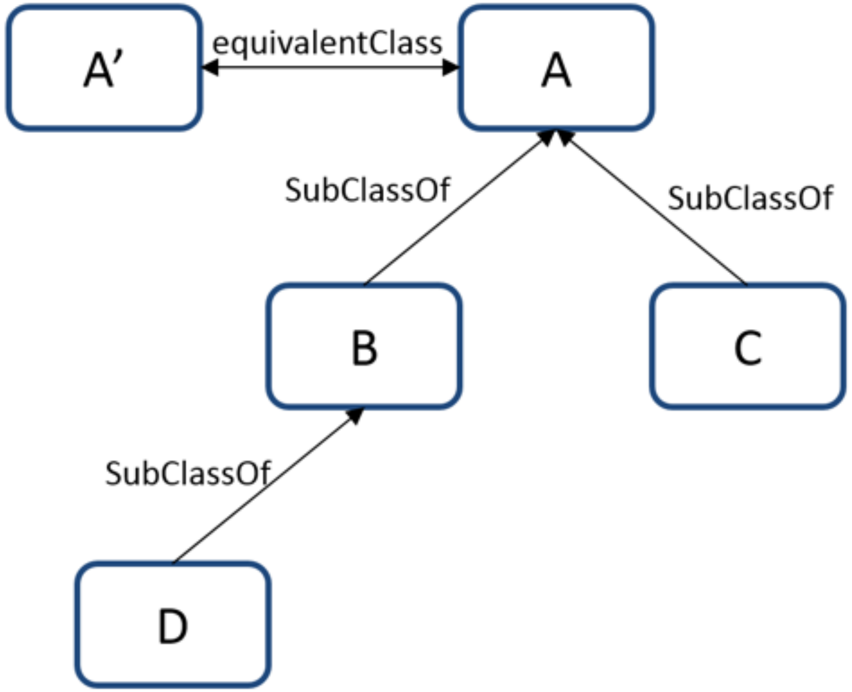
- Create term – data matrix
- TF-IDF normalization, SVD
- Calculate cosine similarity

Similar

Not

	Data 1	Data 2	…	Data n
Query 1	5	5	…	1
Query 2	10	10	…	0
Query 3	0	1	…	9
…	…	…	…	…
Query m	3	2	…	2

Existing ontology (SWEET)



$$sim(X \rightarrow Y) = \frac{e}{Dist(X \rightarrow Y) + e} \quad (9)$$

$$Dist(X \rightarrow Y) = \sum_{i} Edge(Type_i) \quad (10)$$

Where e is a constant used to adjust the final similarity, $Dist(X \rightarrow Y)$ is the distance from X to Y, and $Edge(Type)$ is a function: if the relation type is "SubClassOf", it returns 1; if the relation type is "equivalentClass", it returns 0; if the relation type does not exist, it returns infinity. The resulting value ranges from 0 meaning no relation, to 1 meaning exactly the same.

Query representation

- Query is represented by using dataset as feature:

$$query_i = (\boxed{N_{dataset_1}}, N_{dataset_2}, \dots, N_{dataset_n})$$

$$\boxed{(N_{view} + m * N_{download})\ (m \geq 1)}$$

N_{view} is the number of times $dataset_1$ get viewed after searching $keyword_i$

$N_{download}$ is the number of times $dataset_1$ get downloaded after searching $keyword_i$

m: download >= view

Normalization factor

- How often does a dataset get viewed in all the searching behaviors? The more often, the more general.

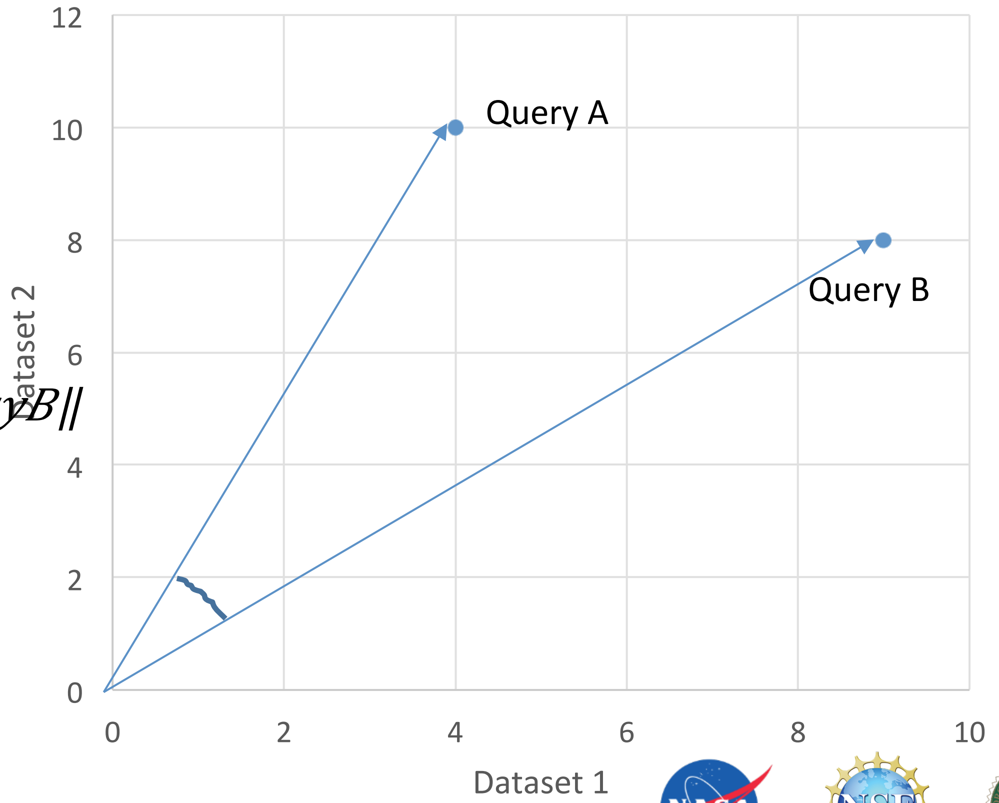
- Datasets related to many keywords < more-specific datasets

$$idf(d_1) = 1 + \log\left(\frac{\text{numKeyword}}{\text{queryFreq} + 1}\right)$$

Cosine Similarity

- Widely used in text mining
- Different from Euclidean dist
- Only orientation matters
- Magnitude is ignored

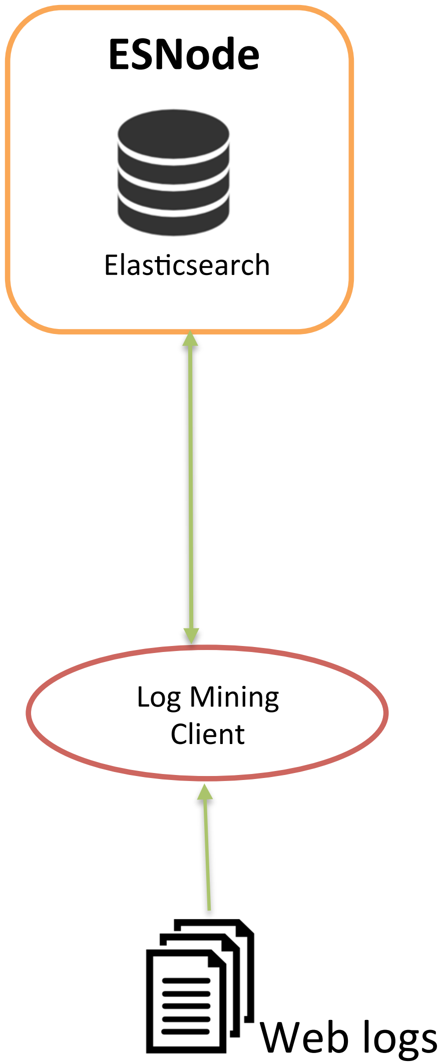
$$sim(A, B) = queryA \cdot queryB \, / \, ||queryA|| \, ||queryB||$$



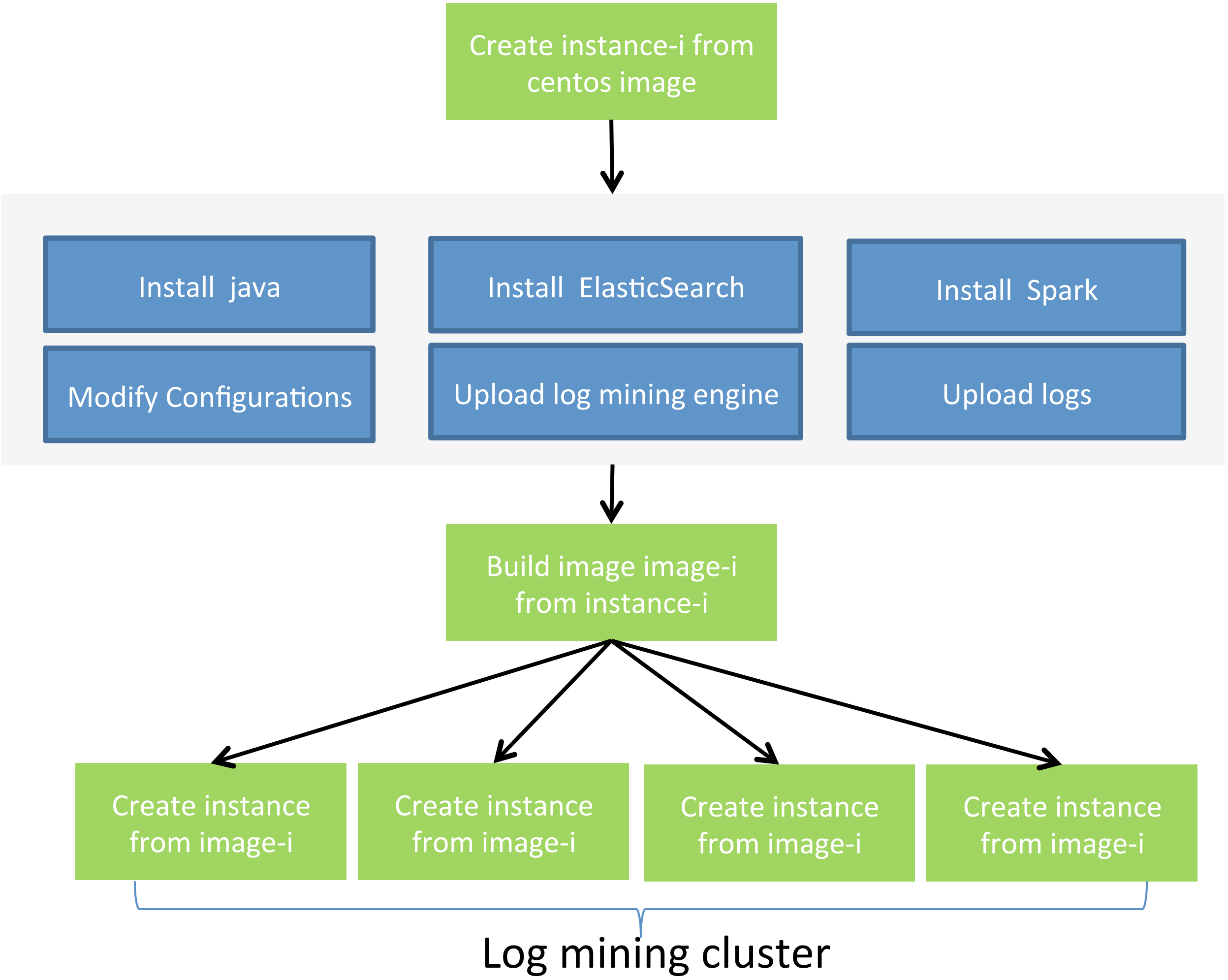
Experiment1 - one client and one node

CPU Core	Memory (GB)	Elasticsearch
4	16	1.7.2

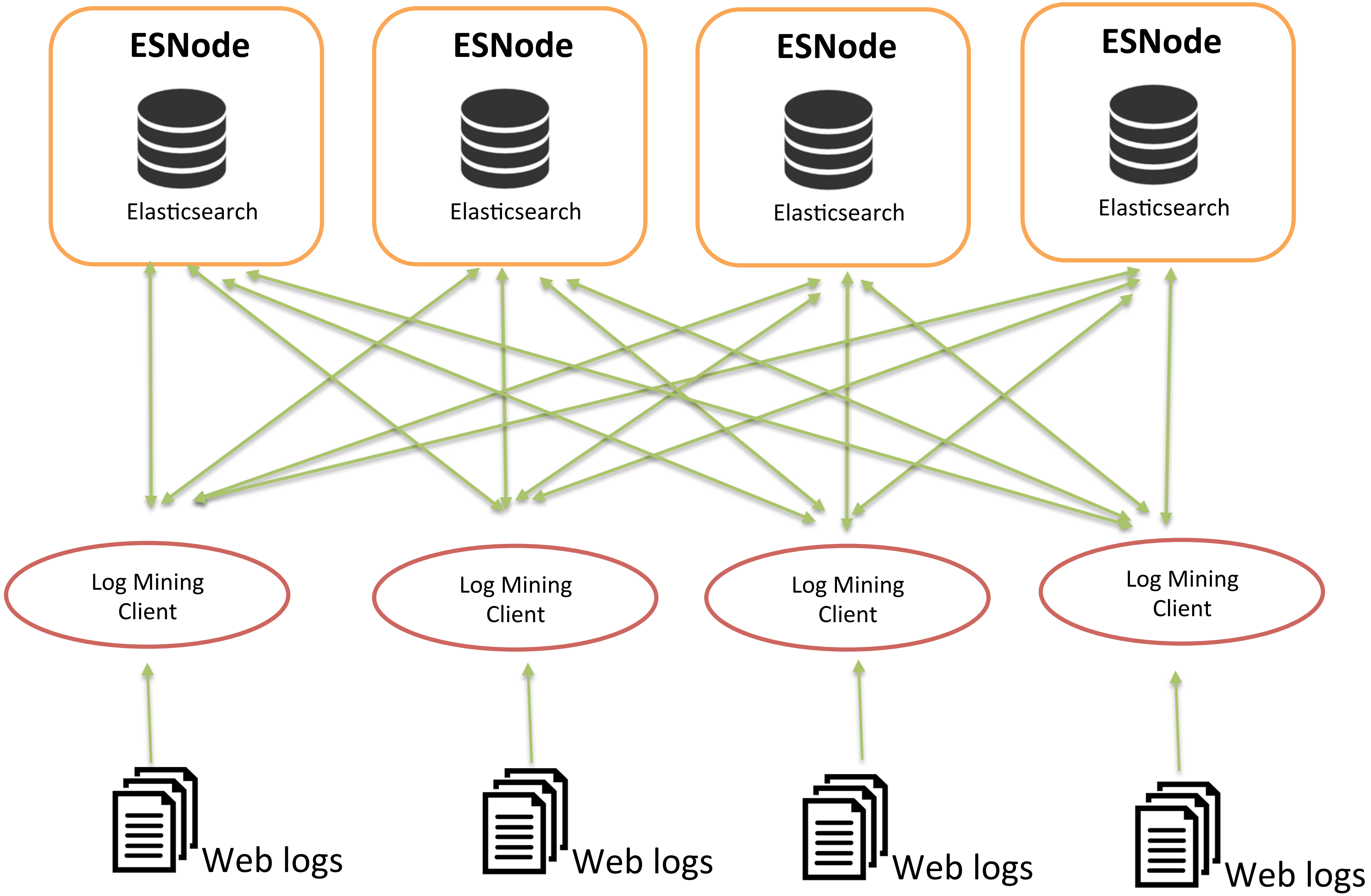
Month	Processing Time(S) (One node)
201401	1387
201402	1287
201403	541
201404	1110
201405	1020
201406	943
201407	983
201408	830
201409	745
201410	751
201411	977
201412	1087
Total	11611(3hours and 13 minutes)



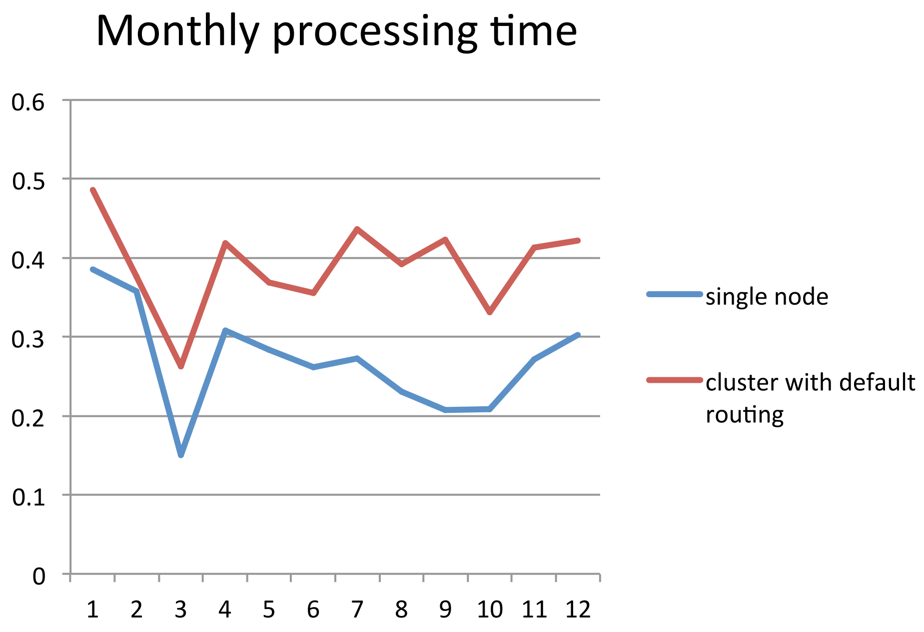
Leverage AIST cloud to set up a cluster

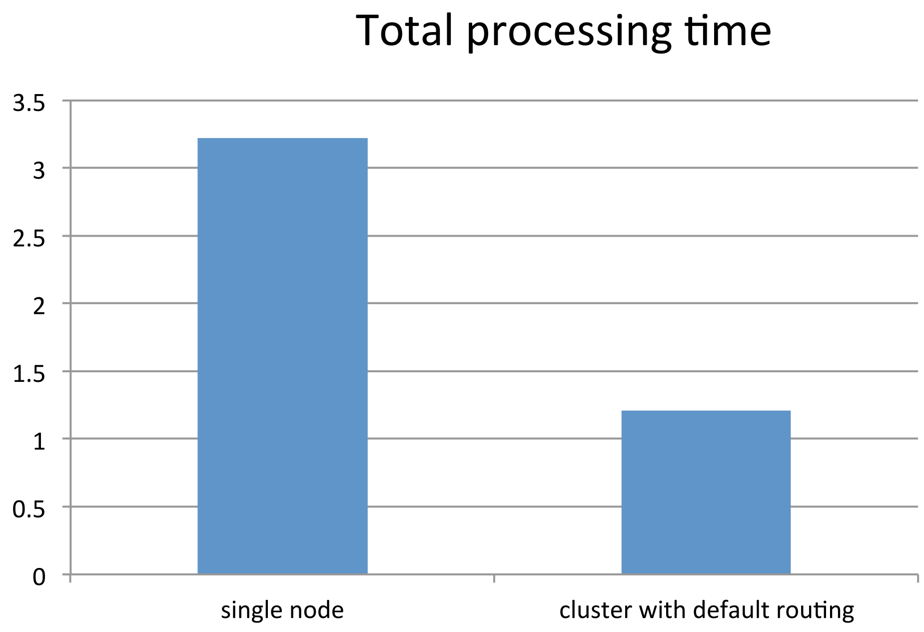


Experiment2-cluster with default routing



Testing Results

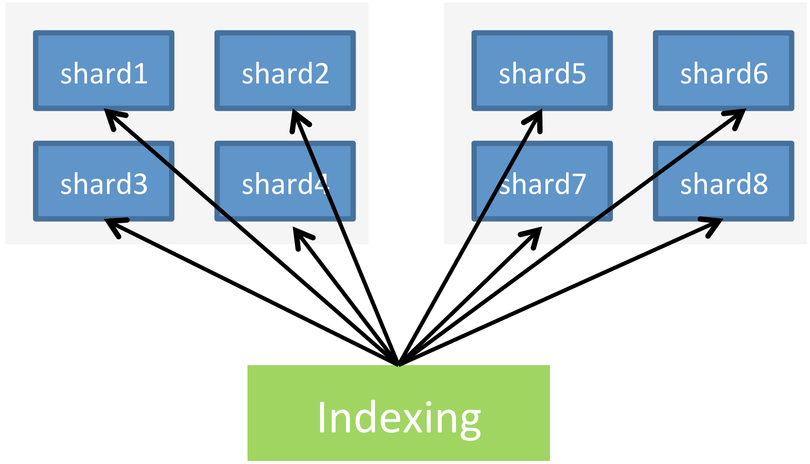




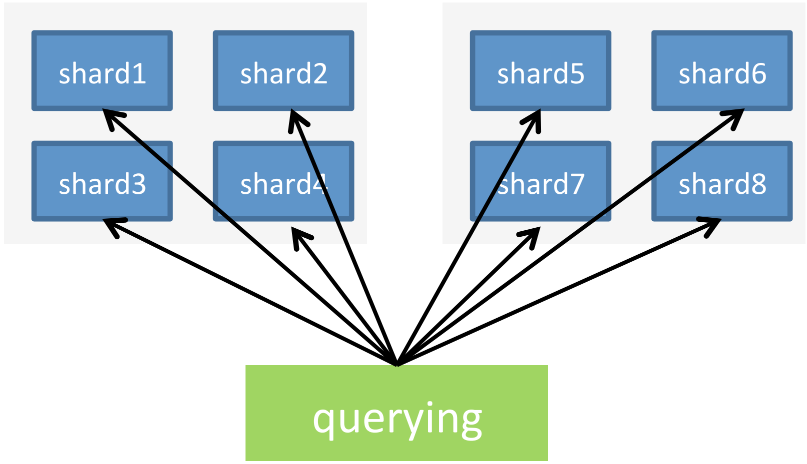
1. Due to network transportation cost, the processing time of using cluster is longer than using single node to process each month's log.
2. The total processing time of cluster is much shorter than that of one node.

Tag & Routing

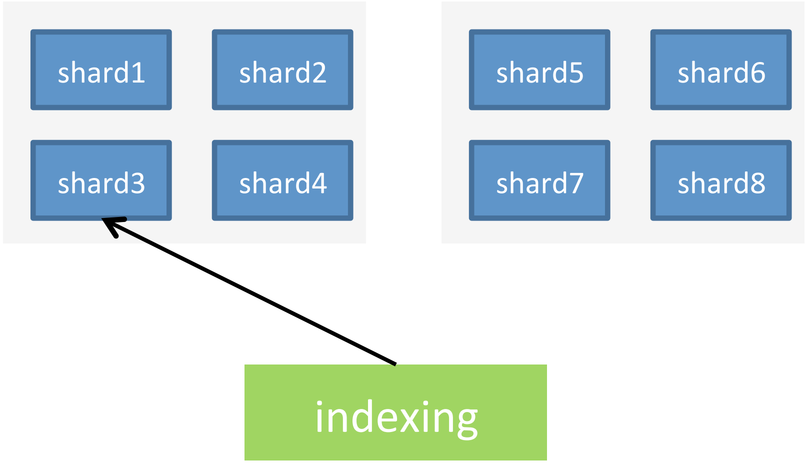
Indexing without tag & routing



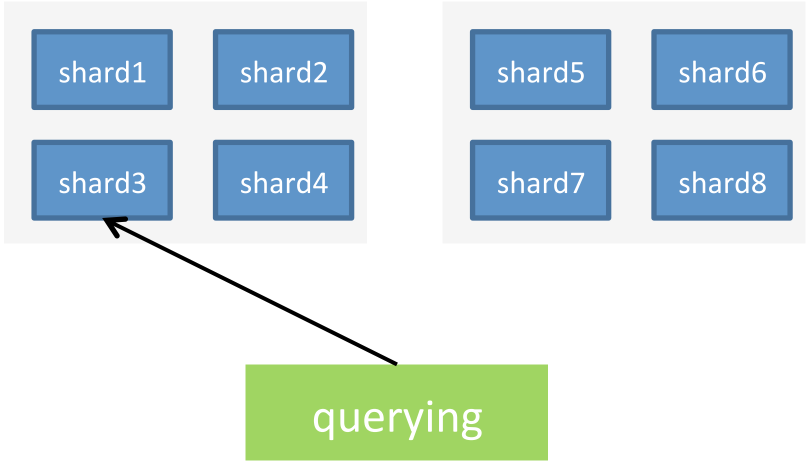
Querying without tag & routing



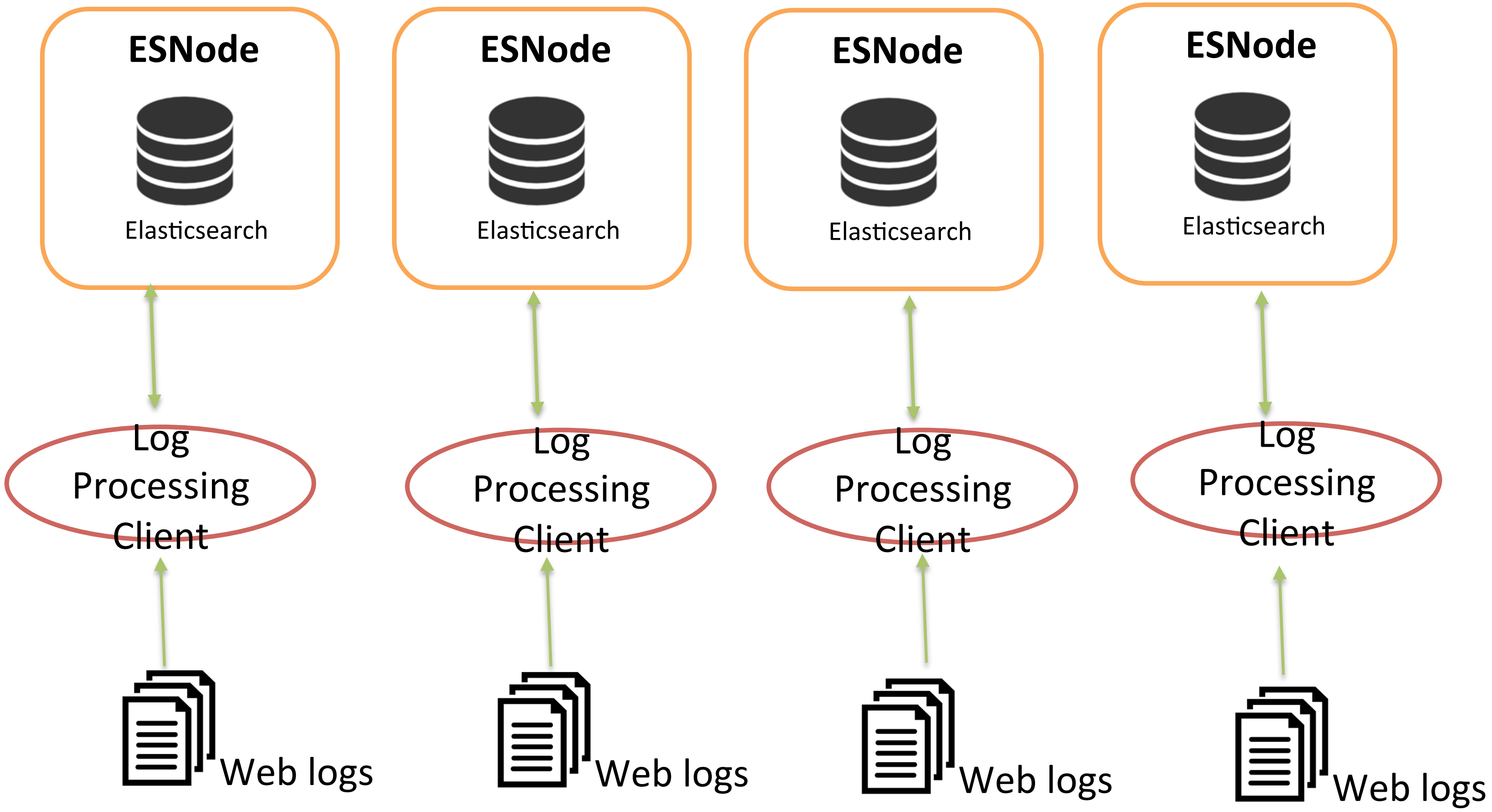
Indexing with tag & routing



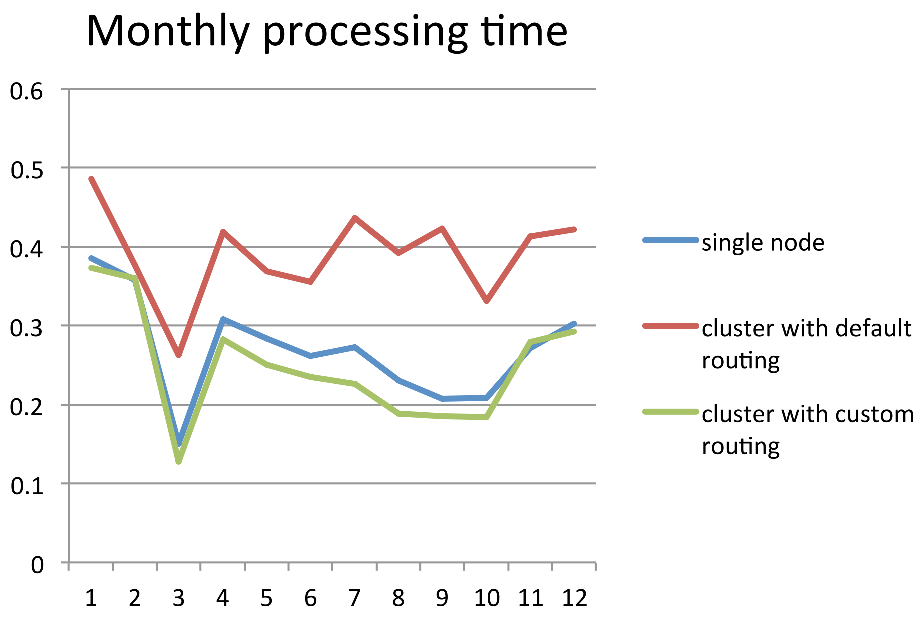
Querying with tag & routing

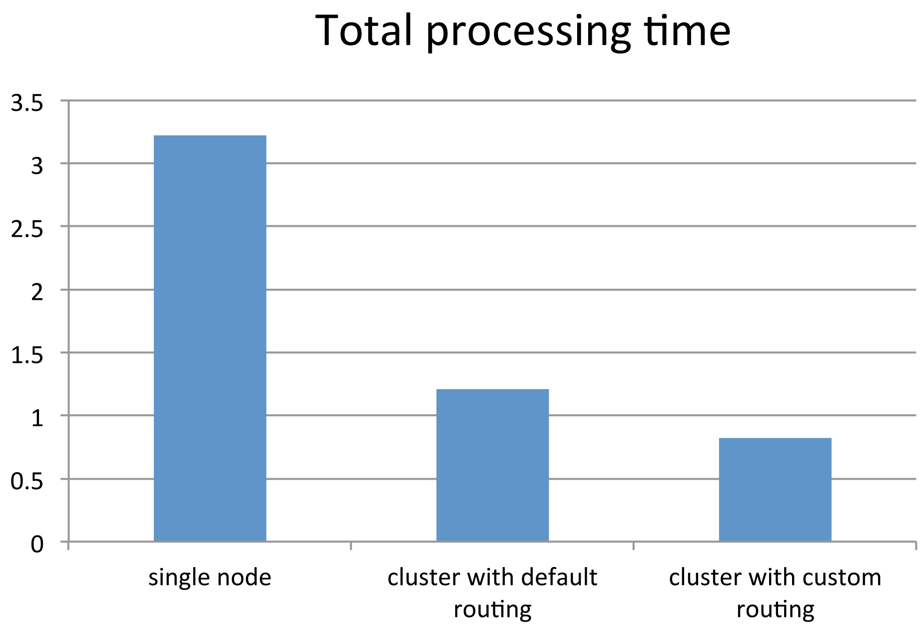


Experiment3-cluster with routing



Testing Results





1. The monthly processing time of using cluster with custom routing is much shorter than using cluster with default routing.
2. The monthly processing time of using cluster with custom routing is equal or shorter than using single node.
3. The total processing time of using cluster with custom routing is shortest.

Ranking algorithm

- Current ranking score is calculated based on Lucene Practical Scoring Function

$$score(q,d) = coord(q,d) \cdot queryNorm(q) \cdot \sum_{t \text{ in } q} \left(tf(t \text{ in } d) \cdot idf(t)^2 \cdot t.getBoost() \cdot norm(t,d) \right)$$

- Factors considered in this formula
  - *Term frequency(tf)*: the more frequent the query is in a certain doc, the more relevant the doc
  - *Inverse doc frequency(idf)*: how often does the term appear in all documents in the collection? The more often, the lower the weight
  - *Field length*: the shorter the field where query appears, the more relevant (e.g. the title has higher weight than abstract)
  - *Query boost*: the importance of each sub-query
  - *Coordination factor*: the percentage of query terms appear in the doc

```
"should": [
  {
    "match_phrase": {
      "name": {
        "query": "ocean wind",
        "boost": 1
      }
    }
  },
  {
    "match_phrase": {
      "name": {
        "query": "surface wind",
        "boost": 0.9
      }
    }
  }
]
```

- The algorithm we are working on will incorporate
  - Query, time-dependent popularity
  - Release date
  - Etc.