CSCI 572 HOMEWORK 2 REPORT

CHANDRASHEKAR CHIMBILI

chimbili@usc.edu

JAYAVANTH SHENOY

jshenoy@usc.edu

MADAN PATIL

madanpat@usc.edu

SREEPADA RAO SINGEETAM

singeeth@usc.edu

Task 1: Hangout session with Dr. Burgess

In the hangout session, Dr. Burgess spoke more on reduction on sea ice and reasons for it.
Following were the important points we could derive from conversation with Dr. Burgess
1) Relation between oil extraction, natural gas reserves and high temperature
2) Sea level affecting islands and life of polar bears
3) Relation between melting and mineral extraction?
4) Where are natural resources concentrated?
5) Find regions where more data is being collected? Why is that?

Speaking to Dr. Burgess helped us understand that co-ordinates, science keywords and location information present in the documents is very helpful.
Dr. Burgess stressed on importance of temperature change and its effect on sea ice. We decided that this should be one of the queries.
We found that energy absorbed by sea ice from Sun is measured watts per square meter (wpm2) and it's proportional to sea temperature.Any document containing this shall be related to the study of temperature and sea ice.
We also set ought to find regions of security interests. These regions were regions with natural resources. Hence one query was dedicated to finding regions with oil, gas, diamonds and rare isotopes.
We also wanted to know if the discussion is all concentrated in few geographical locations. This would help us know if resources are spread out or concentrated.
Since Dr. Burgess said that one important thing scientists could look at was effect of increase in temperature on life at Arctic, we set aside one query for finding out high temperature, reduced ice and their effect on Polar Bears.

Task 3b:

The metadata extracted by SolrCell is minimal compared to what Tika extracts.

1. Nutch + Tika automatically extracts anchor tags from files. This metadata tag is essential for topic identification. SolrCell doesn't extract this by default.
2. It's possible for Nutch + Tika to extract URL from the crawl data whereas SolrCell can't extract URL since we are just using *nutch dump.*
3. Nutch + Tika also has the advantage of building a graph for itself so that it can perform link analysis. URLs extracted from Tika can be used to build graphs that can score based on domains and links.

Ranking Algorithms:

Link based algorithm

We implemented a modified version of PageRank as Link based algorithm.
Our algorithm constructed link graph as follows
1) Read all the URLs from Nutch using Sequence File Reader
2) Get the inlinks and outlinks for each URLs
3) Get the science key words from each URLs
4) Construct a weighted graph with these features
5) Calculate PageRank for the graph

We constructed the weighted graph using inlinks and outlinks with weight as 1.
We also added edge between two URLs if at least 50% of science keywords matched (with weight 0.5).
We also added link between two URLs if they were from same location (calculated using CLAVIN)

The web graph was fed to python-networkx which generated PageRank for each URL.
This PageRank was injected into "page_rank_d" field in Solr.

Content based algorithm

We used tf-idf for content based algorithm. We not only took each word but also considered bigrams.
The bigrams helped us find many keywords like Chuckchi Sea. The text content included title, anchor text, description and many others.
The words in title and description were given more importance using function query in solr (1.5 times as compared to other words).
Also if the query matched science keywords, it was given higher importance (3 times other words).

Effectiveness of Link based and Content based

In our implementation, content based algorithm returned better results compared to link based algorithms.
The content based algorithm ranked the most relevant documents higher.
We found cases where PageRank was high, but was not as relevant as results of content based relevancy.
For example,
Following query "http://localhost:8983/solr/select?q=text:arctic and ((text:"oil exploration"~10) or (text:"iron") or (text:oil))"

Top 2 results of PageRank

Description - Logger records from the Networked Info-mechanical Systems (NIMS), Transect length: ~50m The data was recorded using a CR3000 logger. The sensor trolley was equipped with instruments for recording the distance to vegetation canopy (SR50a Sonic Distance, Campbell Scientific), up- and downwelling short- and longwave radiation (CNR4 net radiometer, Kipp & Zonen), air temperature and surface temperature (SI-111 IR radiometer, Apogee Instruments Inc.) and spectral reflection (Jaz Combo-2, Ocean Optics; GreenSeeker RT100 (505), NTech).
Title - Atqasuk Logger Data NIMS 2014

Description - Spectro radiometer data from the Ocean Optics Jaz Combo-2. Channel 0 (denoted _00jaz in file names) is the fiber looking at the earth's surface, Channel 1 (denoted _01jaz in file names) is the fiber looking at the reference.
Title - Toolik Spectrometer Data 2014 - Ocean Optics, Jaz

Top2 results of content based

Title: Project: Collaborative Research: Diamonds and Oil from the Tundra: A System Study on the Impact of Changing Seasons on Mining and Oil Exploration

Title : Collaborative Research: Diamonds and Oil from the Tundra: A System Study on the Impact of Changing Seasons on Mining and Oil Exploration NSF Award 0902130

Note that content based returned better results as compared to link based.

Testing

Content-based:
We compared our java-based tf-idf output with Solr's default tf-idf query output and we saw that both of them were mostly similar.

Link-based:
Since we are using networkx library to calculate PageRank, the PageRank is calculated accurately. Testing the code that builds the graph is pretty challenging because we are using many factors such as links, keywords and locations and there is no such scoring mechanism in Solr. So, we tested the top k query results and saw that all of them had more inlinks and keywords than the rest.

Indexing Process:

Nutch + Tika + SolrIndexing was easier because Nutch provides Solr with all the required information such as links, metadata, content etc. which is essential to determine the relevancy of documents. Using Nutch + Tika + SolrIndexing eliminates the need to obtain the data by making API calls and parsing the dumped files.
The problem with using just SolrCell is not only the ones described above, it's also inefficient, and delivers inferior results.

Task 5:

Queries: We have listed the queries performed and the information that can be inferred from the result.
1. What time-based trends exist for discussion of oil, iron and other natural resources in the Arctic region? Are documents and topics collocated to geographic region?

a) Indeed, there exists trend in the collection and discussion of oil and other resources. We found pages every year right from 1966 till 2014. This shows that people were interested in oil and resources as old as 1962 and as recently as 2014

['1962, '1966, '1967, '1969, '1970, '1971, '1972, '1973, '1974, '1975, '1976, '1978, '1979, '1980, '1981, '1982, '1984, '1985, '1986, '1987, '1988, '1989, '1990, '1991, '1992, '1993, '1994, '1995, '1996, '1997, '1998, '1999, '2000, '2001, '2002, '2003, '2004, '2005, '2006, '2007, '2008, '2009, '2010, '2011, '2012, '2013, '2014']

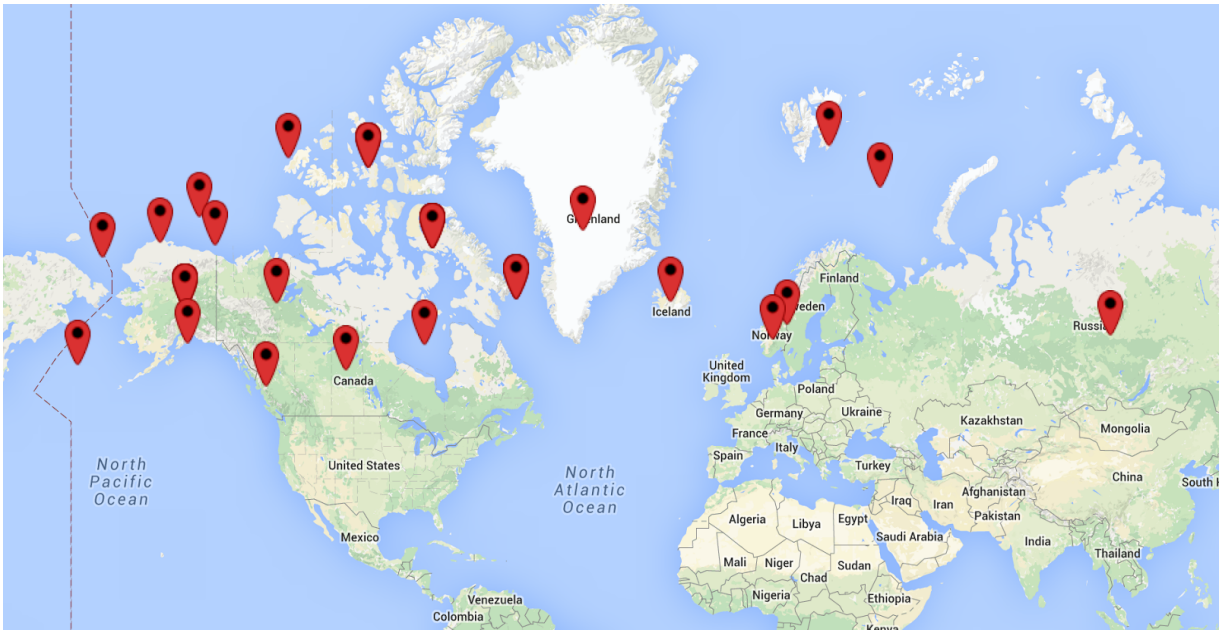
b)Yes these are limited to certain geographic regions mainly Canada, Northern Europe, Arctic Ocean, Scandanavia, Alaska, Chucki Sea, Russia, Eurasia, Hudson Bay

2. How many regions of interest are represented by the data you collected? Identify geographic "regions" as e.g., Circumpolar Arctic region, Antarctica and the Southern Ocean. Can you use the distribution of your documents to represent territories?

There are about 88 regions of interest. Some of them are:

Canada, Alaska, Atlantic Ocean, Alaska Arctic Ocean, Beaufort Sea Vertical Location, Northwest Territories Canada, Geographic Region, Western Hemisphere Canada, Northern Europe, Arctic Arctic Ocean, Bering Sea Arctic Ocean, Arctic United States Of America, Beaufort Sea, Bering Sea, Alaska Vertical Location, Land Surface, Continent, Svalbard And Jan Mayen, Hudson Bay Arctic Ocean, Polar United States Of America, Europe, Land Surface Arctic Ocean, Chukchi Sea.

Here is a plot of all the interest areas:



3. Can you predict areas in which there are national security interests (maritime/air/sea and land)? Which areas and why?

Following are the regions of national security -

Canada, Vertical Location, Eurasia Geographic Region, Chukchi Sea Arctic Ocean Arctic Ocean, Beaufort Sea Arctic Ocean, Northwest Territories Arctic Ocean Canada, Alaska Atlantic Ocean, United States Of America, Svalbard And Jan Mayen, Alaska Arctic Ocean, North Pacific Ocean, Norway, Polar United States Of America, Hudson Bay, Chukchi Sea, Arctic North America, Hudson Bay Geographic Region,

All these areas have either oil, natural gas, iron, diamonds, rare isotope elements. However, national security interests are not only because of these resources, but also to study vegetation change, temperature change, study of sea ice cover and weather conditions.

Here is a plot of all the interest areas:

4. Is there a trend with respect to science data and measurements related to Climate Change? Is it time-based and/or geographic region based? What areas show a high document relevancy for sea-ice extent and decline?

We found lot of scientific measurements are related to climate change are both time based and geographic region based. Recent past has shown lot of study in Tundra region - where subsoil is permanently frozen. Lot of study is undergoing, measuring weekly surface temperature, sea ice concentration and ocean heat content.

Note that query-4e is mainly concentrating on measuring how global temprature change is affecting sea ice extent(like ITP- Ice tethered Profiling). Note that focus on the ice extent measurement are those areas with greatest ice retreat i.e., the northern Beaufort, Chukchi, East Siberian, and Laptev Seas.

5. How is sea ice reduction affecting Polar bears in the Polar regions

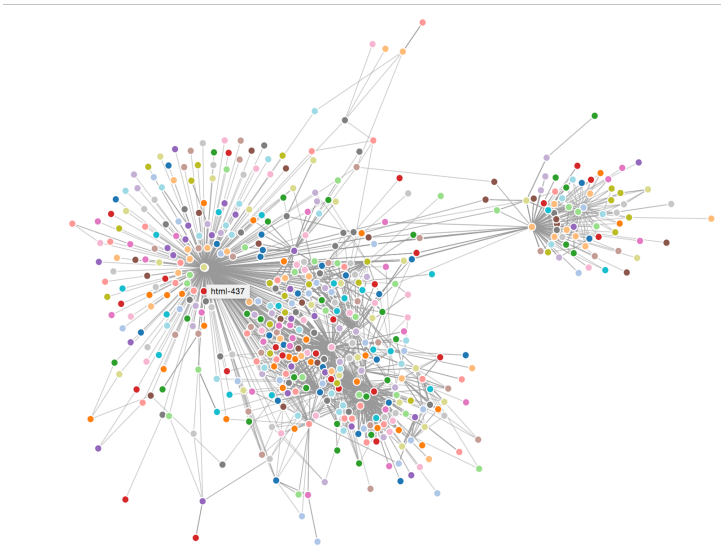
We found that there is research going on to understand reduced sea ice in artic impact polar bears. Note that these studies were restricted to Alaska.

6. In what regions are maritime explorations carried out in Arctic regions?

Eastern Beaufort Sea, Amundsen Gulf, Thule and Summit Station, Mt. Hunter summit plateau, Barrel site near Camp Century, Summit Station, Greenland

Extra Credit

1) We developed a nutch scoring filter. We extended the ScoringFilter interface and added science keywords and location information into Metadata. Scores were based on inlinks, outlinks and number of science keywords matched.
2) We created a d3 based visualization of the link graph from our link based relevancy algorithm. You can find index.html in d3/ folder. Open it in a browser and click the nodes to get the document ID.



Overall Indexing experience

Both Nutch+Solr index and SolrCell made indexing files easy. However Nutch + Solr was much easier and the indexed data was much richer. We had most of the important fields, like co-ordinates, metadata, keywords and so on. But same was not the case with SolrCell. For most of the file types, we had to extract important keywords. By ease of use, we rate nutch+solr better than Solr Cell.