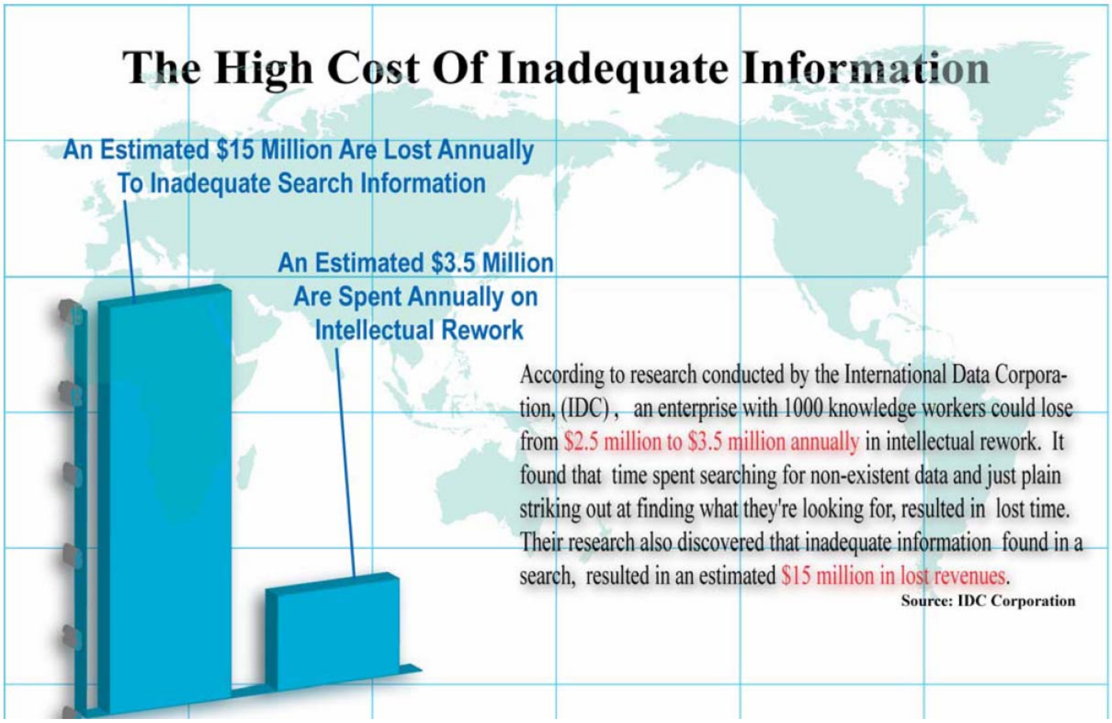
The Search for Unstructured Data
(And the Cost of Not Finding It)

Jess Kozman*
Schlumberger Information Solutions, Houston, TX
jkozman@houston.oilfield.slb.com

Summary

Managers, geoscientists, and new venture team members often need to perform searches for public Web information to retrieve pertinent information for activities such as new area reconnaissance, competitive analysis, environmental awareness, or geopolitical understanding. Performing these searches using standard text keyword searches on public domain search engines can yield non-targeted and irrelevant results. The problem is not lack of information. The problem is finding the right piece buried in the enormous and expanding digital universe of feeds, documents, and web pages located both within and outside the organization. For oil and gas exploration organizations, geography is a primary element of their work and their information taxonomies. A web-based service that offers a simple and efficient way to discover Web based energy-related intelligence can supplement the data generally used to make critical business decisions.



According to research conducted by the International Data Corporation, (IDC) , an enterprise with 1000 knowledge workers could lose from $2.5 million to $3.5 million annually in intellectual rework. It found that time spent searching for non-existent data and just plain striking out at finding what they're looking for, resulted in lost time. Their research also discovered that inadequate information found in a search, resulted in an estimated $15 million in lost revenues.

Source: IDC Corporation

Figure1: Potential cost of inadequate search techniques

Introduction

Traditional search engines fail at these tasks because they have no or limited capabilities to identify geographic locations. They can’t find documents that pertain to, for instance, Australia, British Columbia, or Venezuela because those exact words may not appear in them. Instead, a relevant document may refer only to a region in Australia or a city in British Columbia. Technology is now available to map documents to geography. Through Natural Language Processing, geographic entities can be identified (whether they be oil wells, fields, basins, or lease blocks) within documents and assigned geographic coordinates. The content of more than 3,000 sites of industry targeted information in a Petroleum-specific index can serve as a critical resource for energy industry researchers and analysts. The value of the index is that it enables users to search industry-specific content in a manner that is geographically-relevant to their interests and needs.

Theory and/or Method

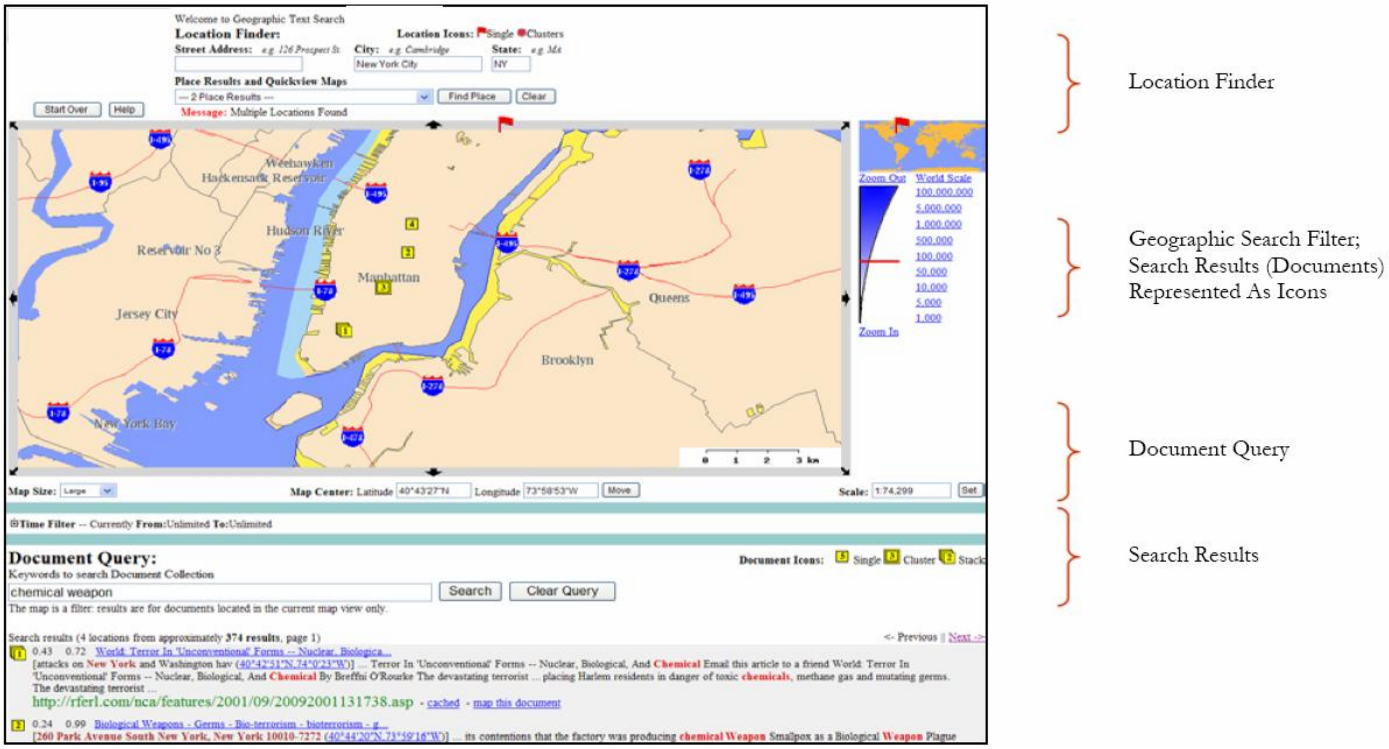
The solution requires a comprehensive geographic text search solution that enables users to rapidly locate high relevance documents by combining text search with geographic and temporal factors, and a production-level geographic entity extractor that parses documents and identifies geographic references within the content. The geographic references are assigned latitude and longitude coordinates and country code tags in an XML file, which may be used as metadata and for processing by third-party systems. An index of Web sites that have been identified as containing content relevant to the energy industry is searchable using map-based Web interfaces. The index can be centrally hosted and easily accessed via aggregated-search features. The source Web sites include those belonging to many organizations in a variety of categories, including:

- Energy companies
- Energy industry associations and trade groups
- Market analysis and business intelligence companies
- Energy industry news sites
- Industry suppliers and services companies
- Energy data vendors
- Consulting companies
- Universities and colleges
- Government agencies

The content sources are continuously updated, and results utilize Geographic Data Modules (GDMs) to identify and locate geographic references. Each GDM contains large datasets of placenames, along with their latitude and longitude, additional datasets, and Natural Language Processing logic. The base GDM contains millions of placenames and other forms of geographic notation. An Energy GDM contains industry-specific taxonomies, such as basins, field names, and operating areas.

Examples

Search queries executed against the indexed datastore will produce results identical in form to other similar web-based queries. When these collections are searched simultaneously with internal collections, the results will be combined with the results from those other collections in order of relevance. Users that rely on the standard graphical user interfaces (GUIs) will receive a list of relevant search results. Each result will include an excerpt from the identified source document or Web site, a GeoConfidence score, a query relevance score and a hyperlink to the source content. Users will also be presented a map showing the geographic region to which the search results relate. Overlaid on the map will be one or more icons, each of which is located at a place that is referred to in one or more of the sources referred to in the search results. At times, results may be provided to sources that require a simple registration or possibly even a paid subscription for copyrighted or commercial data. It then becomes the sole responsibility of users to arrange for access to the sources. The system identifies implied and explicit references to geographic locations within documents, assigns latitude/longitude coordinates to the references, indexes the document, and then enables a search for indexed documents through Graphical User Interfaces (GUIs). A web-based interface or ESRI ArcMap interface can be used.



Welcome to Geographic Text Search

Location Finder:

Location Icons: Single Clusters

Street Address: *e.g. 126 Prospect St.* City: *e.g. Cambridge* New York City State: *e.g. MA* NY

Place Results and Quickview Maps: --- 2 Place Results ---

Message: Multiple Locations Found

Map Size: Large Map Center: Latitude 40°43'27"N Longitude 73°58'53"W Scale: 1:74,299

⊞Time Filter -- Currently From:Unlimited To:Unlimited

Document Query:

Keywords to search Document Collection

chemical weapon

The map is a filter: results are for documents located in the current map view only.

Search results (4 locations from approximately **374 results**, page 1)

1 0.43 0.72 World: Terror In 'Unconventional' Forms -- Nuclear, Biologica...
[attacks on New York and Washington hav (40°42'51"N,74°0'23"W)] ... Terror In 'Unconventional' Forms -- Nuclear, Biological, And Chemical Email this article to a friend World: Terror In 'Unconventional' Forms -- Nuclear, Biological, And Chemical By Breffni O'Rourke The devastating terrorist ... placing Harlem residents in danger of toxic chemicals, methane gas and mutating germs. The devastating terrorist ...
http://rferl.com/nca/features/2001/09/20092001131738.asp - cached - map this document

2 0.24 0.99 Biological Weapons - Germs - Bio-terrorism - bioterrorism - g...
[260 Park Avenue South New York, New York 10010-7272 (40°44'20"N,73°59'16"W)] ... its contentions that the factory was producing chemical Weapon Smallpox as a Biological Weapon Plague

Figure2: Web-based user interface

Results from a search query appear as icons on a digital map and in a results list. The location of each document icon coincides with the geographic locations mentioned within the document. If a user wants to find all documents relative to a geographic area – a country, city, LAT/LON pair or U.S. street address – the system renders a map appropriately speckled with icons representing every document that includes text pertaining to the identified location. By clicking on a document icon, a user gains direct access to the original document. The Web GUI is designed to be an easy to use interface to select the geographic region and keywords used for document searching and to navigate the results both visually on a map and as a list of results with summaries. The GUI is divided up into distinct sections. These sections include the location finder, geographic search filter, document query box, time filter and search results.

The location finder section allows a user to enter the name of a location and set the initial scope of the geographic search filter. Often there are multiple possible locations for a geographic name and the location finder displays the options for user selection. The geographic search filter displays the geographic extent in the form of a map, that bounds the search. This section also allows the user to fine tune the geographic extent by adjusting the scale and zoom level. The document query section contains the query box, where users enter the keyword(s) and optional search operators to be used in the final query. The time filter section is within the document query section and supports the ability to enter date and time information to create a temporal bound for search results.

Conclusions

A unique geographic text search solution provides organizations with a powerful new way to identify geographically relevant information in an efficient and timely manner. This can eliminate hours and days of manual effort tagging geographic references in documents, or searching for material in disparate computing systems that may or may not be ultimately relevant. Instead, analysts can find information that matters.

Acknowledgements

The author acknowledges the technical support of MetaCarta Inc. for material and examples used in the paper.

References

MetaCarta, 2007, Location-Based Intelligence for the Oil Industry:
http://www.metacarta.com/docs/Corporate_White_Paper.pdf

Ventana Research, 2006, Why Geography Matters to the Enterprise. Using Location To Improve Performance:
http://www.ventanaresearch.com/register/login.aspx?req=/primaryresearch/basicdownload.aspx?fid=Why_Geography_Matters_to_the_Enterprise_Using_Location_to_Improve_Performance.pdf

MetaCarta, 2005, A White Paper on MetaCarta's Technology and Products:
http://www.metacarta.com/docs/Corporate_White_Paper.pdf