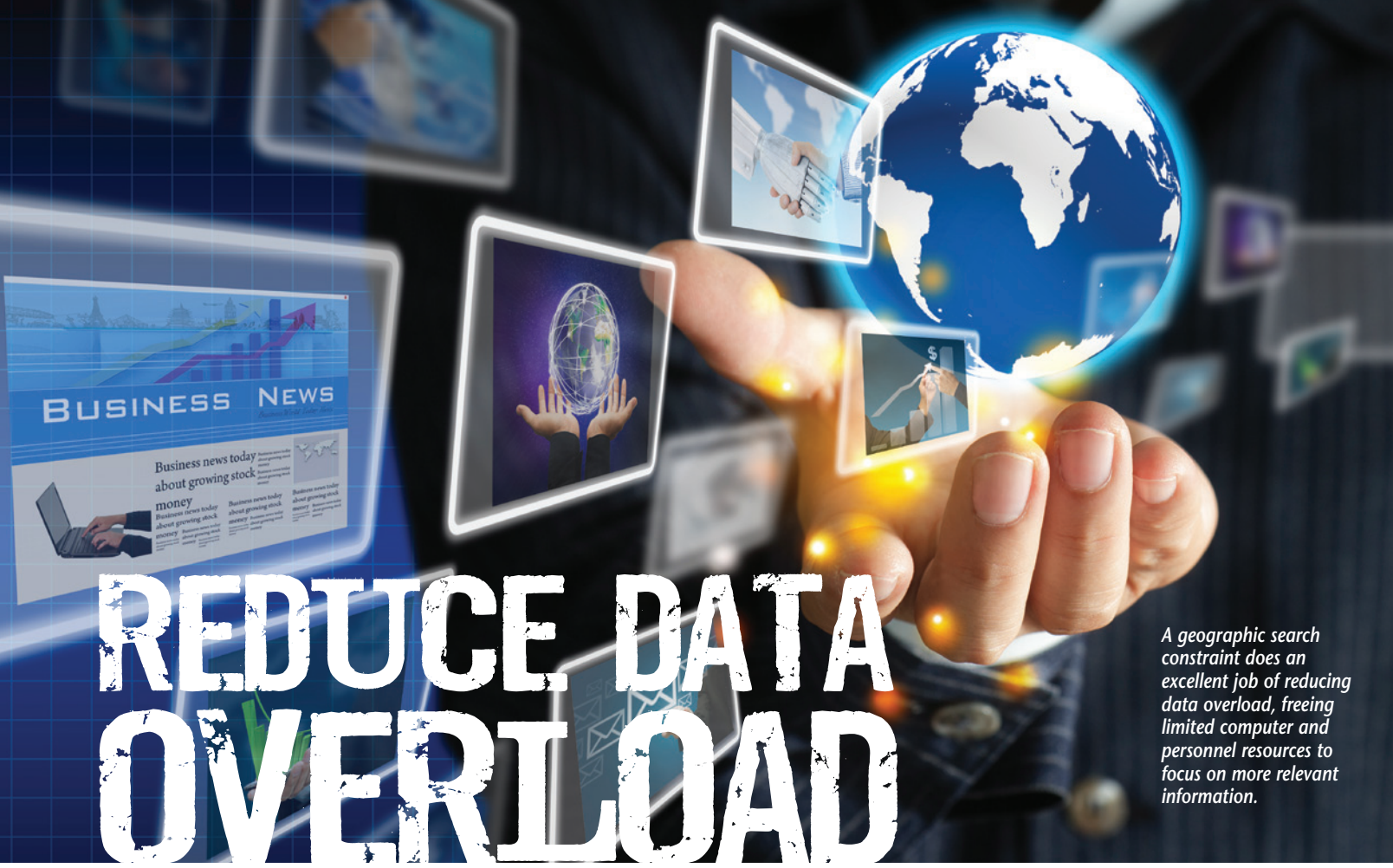


REDUCE DATA OVERLOAD

A geographic search constraint does an excellent job of reducing data overload, freeing limited computer and personnel resources to focus on more relevant information.

DON'T FIGHT MASSIVE FILES.

Discover the advantages of geographically faceted data searches.

*By Dr. Craig A. Knoblock, Geosemble Technologies (*www.geosemble.com*), Manhattan Beach, Calif.*

Data overload is a major problem for organizations, and the problem is getting worse across all industries. Recent IDC research indicates that information workers typically spend a staggering 17.8 hours per week searching for and gathering information. That's a cost of more than $31,000 per worker each year, assuming a $55,000 average employee salary with 30 percent benefits.

At the enterprise level, larger companies and government organizations are wasting vast resources—as much as $5.7 million annually for a 1,000-person organization—searching for and re-creating existing information. Tellingly, IDC notes that "Automating repetitive steps and eliminating those that waste time will increase information worker productivity and save an organization millions of dollars." In this context, let's discuss automating those repetitive steps to save time and resources.

How Does the World Deal with Data Overload?

There are two main tools for reducing data overload through search: topic filtering and time filtering. Both do a good job of reducing data overload, as they eliminate information that's irrelevant to your interests. However, most searches for information carry an unexpressed or under-expressed user assumption: "Limit my results only to the area in which I'm interested."

Whether you're searching near where you are, where you plan to be or within some pre-defined area of responsibility, a geographic search constraint does an excellent job of reducing data overload, freeing limited computer and personnel resources to focus on more relevant information.

In short, with topic and time filtering, an effective geofaceted search capability can be an important contributor to reducing costs and accelerating knowledge in organizations that have areas of geographic responsibility. Given these benefits, it's worth considering some technical approaches for automatically linking textual content to places and compare some best-fit scenarios for the different techniques.

Geographically Faceted Search and Discovery

The National Academy of Sciences estimates that 80 percent of online content contains geographic information—much of it unassociated with any address or latitude/

longitude information. Such online content could contain a vast, unknown number of opportunities and threats, which if found and leveraged in time could be valuable to a range of organizations seeking a competitive advantage in their areas of geographic interest.

As the number of text documents continues to grow, the problem of organizing and searching within documents becomes increasingly important. Search engines, such as Google, can locate documents by keywords, but in many cases a user may not know what search terms to enter to find relevant information or the keywords are too generic, resulting in too many matching documents from every corner of the globe.

One way to address these problems is to use a document's contents to link it to one or more geographic locations. There are several approaches to solving this problem, the most common of which are the natural language processing (NLP) approach to linking geographic references and the term-frequency approach to linking documents to locations.

NLP Approach

The approach taken in most commercial systems today is identifying each geographic reference in a document and determining the geospatial coordinates of that reference. This is the approach taken in widely known systems such as MetaCarta and Yahoo! Placemaker. The general method to perform this linking begins with a technique called spotting, which uses a large gazetteer—a database of place names—to identify every possible geographic reference in a document.

This is a challenging problem, because many terms can have geographic and nongeographic meanings. For example, many common English words, as well as words in other languages, also are place names, such as the cities of To, Myanmar, and Of, Turkey.

This problem can be addressed by performing NLP on a document. After spotting, the NLP system performs part-of-speech tagging, which determines how each term is being used in a sentence and the context in which it is used. Thus, a word being used as a preposition would be unlikely to be a place name and the natural language processing can help resolve these ambiguities. This process requires a set of language-specific rules trained into the system, so the rule set used for English differs from the one used for Burmese, Turkish, etc.

Once the system has identified the geographic terms, the next step is to resolve each of the geographic references to a specific location. The challenge is that any given geographic term can have tens, hundreds or even thousands of places around the world with that name. There are a variety of approaches to solving this problem. A common approach is to maintain statistics on the frequency with which each term refers to a particular place. Then when a term is encountered in a document, the system generally will get the correct assignment by assigning the most likely meaning of the term. Thus, the term "London" usually will be interpreted as London, England, unless there's additional information to indicate otherwise. The problem with this approach is that if "London" was actually referring to "London, Conn.," then the system would be unlikely to find this assignment unless "Connecticut" is explicitly mentioned.

Once the appropriate location is determined for each geographic reference, a system can associate the corresponding latitude and longitude coordinates with each reference. Such systems link every geographic reference, so for any given document one could have hundreds—even thousands—of references linked to locations. Because the systems don't take a position on the geographic focus of a document, even a passing reference to a location will result in a document being linked to that location.

Systems employing the NLP approach will return all of the locations mentioned in a document. In some cases this is sufficient, but in others, the real goal is to determine what's called a document's geographic focus, which is the location or locations that are the document's primary focus. For example, an article might be describing Disney Hall in downtown Los Angeles and mention in passing that the architect, Frank Gehry, also designed other titanium-covered buildings, such as the Guggenheim Museum in Bilboa, Spain, and the Stata Center in Cambridge, Mass. Although these other geographic references can be correctly disambiguated, the issue is that the real focus of the document is Disney Hall in downtown Los Angeles. There has been some recent work on a research system called NewsStand, which attempts to identify the focus of a news article to place it on a world map. It does this by combining the various geographic references in the document to determine the likely overall focus, but it has a limited gazetteer, and the focus only can be determined based on the combination of terms that have been linked individually.

Term frequency-inverse document frequency (TF-IDF) is a way to determine which terms in a document should be weighted most heavily when trying to understand what the document is about. Here a TF-IDF word cloud is based on a corpus of news, blogs and tweets about Syria.

THE NLP-BASED APPROACH

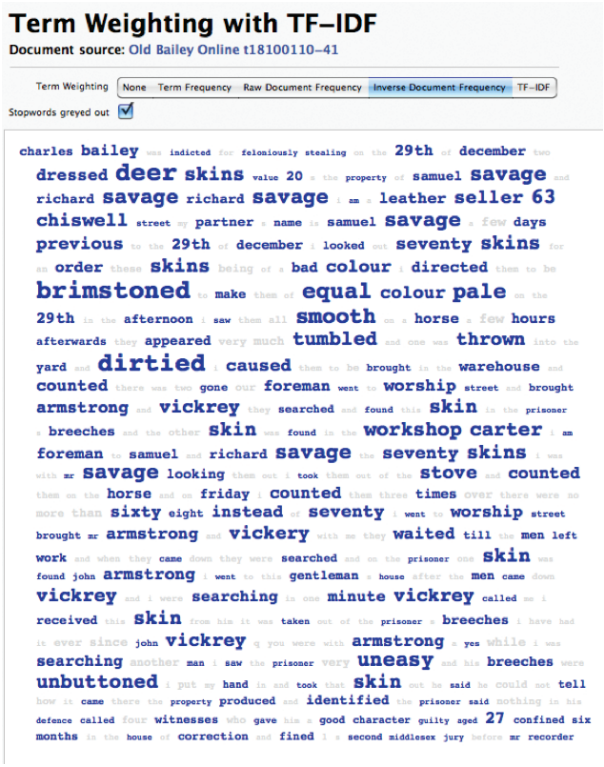
works well with a large repository of documents and an application that requires finding any mention of a geographic location within those documents.

Term-Frequency Approach

Instead of considering each geographic term in a document and identifying the most likely geographic location, an alternative approach is to use the combination of all of the terms in a document to identify the most likely location described in a document. This is similar to the approach used in search engines, which index all of the terms in a document and use the frequency of each term to determine how similar it is to another document.

In contrast with the NLP-based approach, instead of considering each document and each geographic reference in the document, a term-frequency system first identifies the location of interest and then constructs a set of keywords based on the location. Similar to the search engines, the system then can quickly and efficiently find all documents that match the combination of those keywords. This approach is used in Geosemble's GeoXray product to accurately link documents to locations.

Term frequency considers the complete set of terms in a document to compute the most likely location that's the focus of the



A term's weight increases proportionally to the number of times it appears in a document, but it's offset by the term's frequency in a repository of documents. This helps to control for the fact that some words are generally more common than others.

document. When such a system performs this linking, it also computes a corresponding score, which captures the confidence level that a document is about a given location. In some cases a document may have more than one geographic focus, and the system assigns a score to each location.

Because the term-frequency approach doesn't need to separate out geographic references, the system can use other types of information to perform the linking, such as names of businesses, street names, people who work there, phone numbers and other associated information. In general, just because a geographic location is mentioned in a document, the system wouldn't link it to that location. Rather, there would need to be sufficient evidence in the document that the location was a topic of the document.

Another important advantage of the term-frequency approach is the ability to perform fine-grained linking to locations. This means that instead of merely linking documents to a city or general area, the approach can link documents down to specific buildings, individual businesses or even people associated with locations.

The GeoXray product performs this fine-grained linking by using a gazetteer with specialized "place signatures" for a region and then computing the documents that link to each of the individual locations. The fact that the term-frequency approach determines the overall geographic focus for the documents makes this fine-grained linking possible. Otherwise, a system would be overwhelmed with a detailed gazetteer if it tried to link each item mentioned in a gazetteer to an individual location.

For example, consider what would happen using the NLP-based approach if you put "McDonald's Restaurant" in the gazetteer—McDonald's has more than 30,000 locations worldwide! This ability to perform fine-grained linking makes it possible to build applications where linking down to specific buildings or businesses, such as "the McDonald's on Culver Blvd.," is required.

Pros and Cons of Each Approach

Both approaches have advantages and disadvantages. The most applicable approach depends on the details of the specific application.

The NLP-based approach works well with a large repository of documents and an application that requires finding any mention of a geographic location within those documents. Because each reference can be assigned at processing time, it also means the documents can be fully processed in advance and finding the documents that mention a specific location can be performed quickly. The primary disadvantage of this approach is that it focuses on disambiguating only the geographic terms in the gazetteer, and it's difficult to accurately compute the overall geographic or topical focus of a document. In addition, because of the complexity of processing natural language syntax and rules, it's computation heavy, and new rules must be produced and processed for each language.

One disadvantage of the term-frequency approach is that instead of preprocessing all documents and being able to determine every document that mentions a location, the approach must be given the location of interest first, then it finds the relevant documents. However, in practice, most users know their geographic area of interest, potentially mitigating this disadvantage.

THE TERM-FREQUENCY APPROACH combines all of the terms in a document to determine the document's geographic focus.

On the positive side, the term-frequency approach combines all of the terms in a document to determine the document's geographic focus. In addition, this approach can handle a much more fine-grained gazetteer and exploit nongeographic terms, thereby improving the ability to accurately link documents to locations and opening the option to geofacet entities as well as places. And because it's a text-matching technique, term frequency scales well and works in any language.

GeoXray allows users to understand the content landscape and drill into areas of interest. For example, by reviewing content about Syria, relevant topical content is pinned to the actual point of interest.