Advanced Data Capture Solutions

RAF Technology Technical Whitepaper

RAF Searches for relevant data

RAF's semantic search engine finds relevant data in text and quantifies it for later use. This series of Technical whitepapers seeks to reveal the true value of semantic search and to break the process down into simple elements that can be critiqued. This whitepaper will outline several concepts that will define each element of the semantic search process in terms that can be implemented. The goal for any successful semantic search strategy is to provide sufficient refinement within your search process in order to find quantified data for use in modeling and informative data which can be presented to users.

This first paper will address some simple approaches to finding concepts within searched documents. Later papers will concentrate on other details of the process and their implementation.

Introduction of basic concepts

Before we can begin our discussion on semantic search technology, we must first address two issues that are critical to the fundamental understanding of the limitations that are exhibited in current search technologies. In order to do that, we must first define the term information and related terms. Second, we must understand the relationship between information and a document. For the sole purpose of this whitepaper, the following terms are defined as follows:

Data: Words, phrases, and terms without abstract mappings or context.

Concept: A general idea derived from or inferred by specific instances or occurrences of data.

Information: An abstraction of a collection of similar concepts to convey a cohesive generalized idea.

Identifying information is an abstraction process which occurs when generalizing data into concepts and views. A document is an attempt by its author to encompass both data and a representation of concepts. Since humans think more in abstract terms, it is easy to understand why most search engine technologies generally fail to achieve the desirable outcome which is locating information not data. Therefore, the identification of concepts becomes a key element to semantic search.

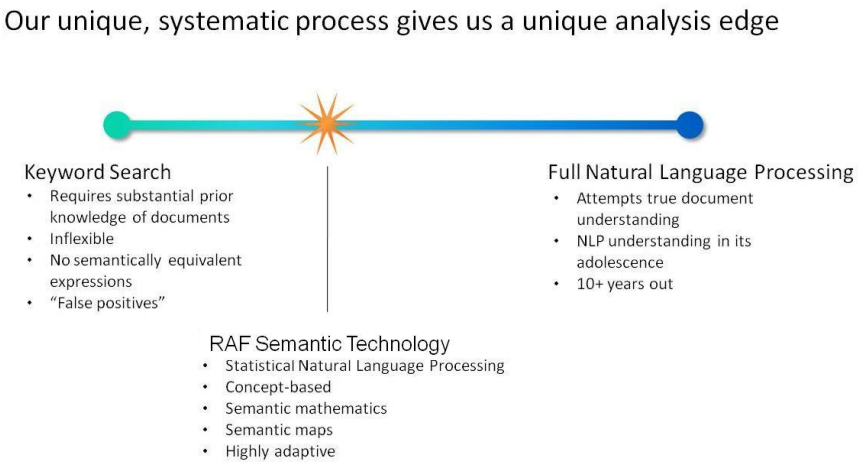
To illustrate this point, let's say you were interested in obtaining a list of financial institutions with deteriorating financial condition. By using conventional search queries, you would probably obtain random data such as financial institutions

market turbulence, or deteriorating cost efficiency. Of course what you really desire is a report summary that lists all failing institutions. The example highlights the shortcoming of current search technologies because today's search engines merely highlight random bits of data devoid of context which are located someplace in a document. Google can search documents and key words, but generally cannot present concepts. In most cases, the extraction of real information will require data to be found and collected from multiple document sources in order to achieve the highest possible information value.

Approximating a Semantic Search

So why is semantics so important? It is important because semantics can identify concepts and thus extract information from data. In other words if you are looking for the meaning of documents, you will need to capture concepts. Long before we deal with extracting or quantifying information within or across documents, we have to find that information. The first step in doing that is to know what we are looking for. Let's start with understanding the basic semantic search unit -- concept. A concept is basic enough that a Boolean search can be written that locates a significant amount of data that is mostly relevant to the concept. Ideally, then, when a document set is indexed, it is broken down into tagged concepts and their interrelationships. A query would likewise consist of a set of concepts and their defined relationships, and a search would find which documents contained concepts and relationships that matched the query. This is, after all, roughly how Boolean search engines use words.

The problem is, of course, that while we are all more or less in agreement on what constitutes a word, there is no agreement at all on what constitutes a semantic representation of a concept, no yet-defined way to present or search for their relationships, and therefore no ideal way to conduct a generalized semantic search for information. Therefore, our first and most basic goal is to approximate the ideas of concepts, of relationships among concepts, and of finding places within documents that contain those concepts with the desired interrelationships. There are multiple techniques for achieving this goal. At one extreme are Boolean searches. At the other, the ideal of localizing concepts within and across documents. RAF has found a way to bridge the gap.



Word groups

Word groups as keys to emphasis

If words are the basic units of Boolean searches, perhaps the first step along the path to a general semantic search engine would be to use word groups. Text that occurs in a document significantly more often that might also appear in similar documents is a clue to the conceptual content of the relevant text. Thus if "higher risk" appears out of proportion, the areas within a document where it appears are probably dealing with the concept of higher risk. Word group analysis is therefore a poor man's semantic analysis, concentrating more on what the text is about than on what the text means.

Two kinds of word groups

Two kinds of word groupings are relevant. The first are groups of words that are capitalized, quoted, or otherwise emphasize by the text itself. Thus, if the text includes the words, "Gone with the Wind", there is some likelihood it is discussing Margaret Mitchell's book. Such highlighting is a reasonable guide to the subject under discussion, but may not be very common in some classes of documents of interest. Care must also be taken with capitalization of names and initial words in sentences.

The other kind of word groupings will probably have greater utility, though it also is not free of problems. At its simplest, this second approach just looks at the probability of single words, two word pairs, three word triplets, and the like in the text and sees whether they occur with significantly higher likelihood than in the average document of the kind under consideration.

Determining whether a word group occurs with unusual frequency requires caution, however. Most words in a short document will occur with a frequency far

above their normal appearance in the language or in the overall class of documents. For example, every word in a 100-word document occurs at least 1% of the time but very few words in a large set of English appear this often[1]. We therefore only want to highlight those words and word groups that truly do appear with unlikely probability.

Once a word group is found, its surroundings must be checked for the presence of negations – things that reverse its meaning. Thus the surroundings of "higher risk" must be searched for the presence of negating items ("no higher risk", "higher risk not expected"). Those with negating terms should be counted as part of a negated word group "higher risk".

Word group candidates

The approach of just finding unusually high frequency word groups is too simple. The primary reason is the presence of stop words[2] that do not contribute significantly to meaning no matter how often they appear. Stop words are usually high-frequency words such as "if", "the", and "and".

Word group members should therefore be words that contribute to the meaning of the phrase in which they appear. There are two approaches to selecting such candidate words – positive and negative.

- **Negative approach:** This approach uses as elements of word groups all words except stop words. Stop words do play a role, however, since when they negate or otherwise significantly change the meaning of a phrase must be taken into account.

- **Positive approach:** This approach creates a list of words from which the word groups are formed. Online financial glossaries[3] give places to start in producing these lists. Another possibility is to go through all the example documents and pull out those words that occur with substantially higher frequency than standard English, remove stop words from the list, and use those that are left as the allowed lexicon.

[1] In general English, only seven words occur more than 1% of the time in the "average" document: the, of, and, to, a, in, that. Only 90 occur more than 0.1%, and 1031 more than 0.01% of the time. This means that any document of less than 10,000 words will likely have many words that, by occurring at all, occur well above their normal likelihood. The word frequency (and even more so, the word pair frequency) measurer will have to be careful because of this to not find spurious "important groupings".

[2] A stop word is a common word that does not generally add to the meaning of a sentence. "An", "is", "but", and "the" are good examples.

[3] Examples include: http://biz.yahoo.com/f/g/, http://www.duke.edu/~charvey/Classes/wpg/bfglosa.htm, http://www.investopedia.com/dictionary/, and http://www.forbes.com/tools/glossary/index.jhtml

If you have a complete list of words, or declare the state of any previously-unseen word, there is no difference between the two approaches. All words fall into one or the other category.

Concept families

Word group families approximate these and are our first approach to them. As word groups are an approximation to concepts, word group families are an approximation of concept families.

Word groups are an approximation to the Boolean search we have defined as the quantum of concepts – the most general idea that can be found relatively unambiguously with a Boolean search, regular expression, or the like.

The Semantic Search Advantage

Word groups as keys to emphasis

There are two key concepts that provide differentiation between standard Boolean searches and semantic searches -- what is the object of the search, and what are the components of the search.

In Boolean searches you are looking for a document that contains words in a certain relationship. Thus the object of the search is a document and the unit of search is a Boolean logical structure of words. That structure may be relatively complex, but it is rigid. While some variation may be allowed, you are essentially looking for documents that exactly match the word structure used to search. The reason you are doing the search is to find needed information that is relatively irrelevant to the search. It is the user's responsibility to guess what words are used to point to the information he needs. Notice the contradiction – the user does not possess the needed information (else why would he be doing the search), but he must guess how the needed information is expressed, and hope it is located in one place in a document.

Most of the levels of abstraction that Boolean search requires of the user are handled automatically in a Semantic search. The work of finding the needed information consists only in determining what concepts are within that information, rather than having to go further and guess what word groupings might have been used to express those concepts. A Semantic search translates concepts into word groups "offline", greatly reducing the guessing game part of search.

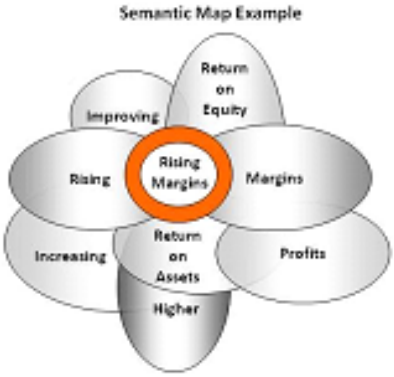
The goal of any search is information, not documents. Since information is an assembly of concepts, a Semantic search is much better aimed at finding what is

really being sought, particularly since the information may not exist in any one document. As the search goal of Boolean searches is documents, so the goal of a Semantic search is information. As the unit of Boolean search is assemblages of individual, so the core element of semantic searches is the Semantic Map.

Semantic Maps

Overview

A Semantic Map translates a concept into different ways of expressing the concept. The following diagram shows a partial Semantic map for the concept "rising margins".



Mental concepts closely related to "rising margins" can be expressed in many ways, most of which use neither the word "rising" nor the word "margins". "Improving return on assets", for example, provides information relevant to "rising margins". Semantic maps translate concepts into the kind things a Boolean search can look for. This is why we refer to a Semantic Search engine as an information compiler which takes concepts used in human thought, and translates them into the "machine code" of individual words and phrases that a Boolean search can use.

There are three important parts to a Semantic search – the creation of Semantic maps, the preparation of document sets for Semantic searches, and the carrying out of Semantic searches. The following sections briefly outline each of these parts.

Creating a Semantic Map

A Semantic map may be created manually or automatically, but map construction consists of determining what concepts are important for a given information universe and then determining the ways each of those concepts may be expressed. A map can be built top-down or bottom up. The above example was constructed top-down. In SEC filings the term "rising margins" is known to be important and discovering whether a company expects its margins to rise is an important task of fundamental analysts. A top-down map is built by interviewing those analysts and asking them how they have found "rising margins" or related information to be expressed in SEC filings. The above map partially shows the results of those interviews.

On the other hand, a map can be constructed from the bottom up. A document set of interest is chosen that has a relatively structured vocabulary and finite concept space. For this application, we are not seeking natural language processing. We are not trying to understand anything that might be expressed in print. Instead, we are attempting three things – determining what concepts are discussed in particular document sets, defining sought information as a set of concepts, and then searching for that information in similar document sets.

Within a reasonably broad subset of the desired document universe, we determine what word groupings occur while keeping in mind such things as highlighting, sentences, and phrase structure. This gives us a broad collection of word groupings, most of which do not express anything important. We then determine which among them are Semantically Improbable Phrases.™ A SIP is a word grouping with two properties: it occurs sufficiently often within the sample document set to matter, and it occurs with a frequency substantially higher than in standard English[4].

Once the SIPs in a document set are determined, their word group components are "flipped" using what is essentially a specialized thesaurus. Each of the words and word subgroups of the SIP are compared with entries in the thesaurus and other SIPs are searched to determine if the flipping has matched one of them. If so, the two SIPS are combined into one Semantic Map, believed to express a single concept. Finally, the user reviews each Semantic Map to determine whether the concept it expresses is unitary, relevant, and unique. A unitary Map expresses only one concept – we haven't accidentally conflated two or more concepts. A non-unitary map is broken apart into unitary maps.

[4] Clearly there is nothing special here about English. This approach will work with any language and would be simpler to implement in, for example, Chinese because of the structure of the language.

A map is relevant if the user believes its concept will be needed in searches. Irrelevant maps are rejected. A concept is complete if it appears in only one map. Maps that express the same concept are combined.

Building the Semantic Maps is the hard part of Semantic searching. Once it is done, applying it to indexing and searching documents is relatively straightforward.

RAF's Technical Approach

The Opportunity

RAF has developed our own Semantic Search Engine technology which is capable of searching multiple source documents that can be used to assemble information in a structured format which provides precise sentiment capture. Our approach offers a significant advantage and benefit because our search technology is not dependent on any one document. Instead we score documents for the relevance of data in order to provide the highest quality of sentiment collection and analysis. This process also provides a user with the ability to reference the source documents that were used to extract the sentiment.

Our approach is very scalable because customers can create their own custom thesaurus which is used to extrapolate subject sentiment and concepts (e.g. Legal Thesaurus) which are unique to their own application.

A search can easily be initiated from a user interface menu by simply selecting predefined semantic maps which were created specifically for your application. Users can also create their own custom search queries.

The output of searchable information is generated to a user level using a web browser interface which enables human intervention and review.

Our Search Methodology

Documents that will later be searched[5] are indexed by the concepts they contain. While *information* may exist only across many documents, *concepts* are sufficiently localizable that documents can be indexed as to whether they contain them. This approach is similar to the indexing that document preparation entails prior to more standard searches. Indexing not only consists of determining what concepts are within a document (and where) but also determining whether they are expressed positively or negatively and with what strength. For example, concepts that contain specific data ("fourth quarter earnings are expected to be up 14%"), the content of that data is indexed along with the concept.

[5] This can also be done in real time, but slows down the search process.

Searching

A Semantic search has two parts to it – creating the query and executing it. A query is a linked set of concepts that express the desired information. Thus the concepts for "analyst consensus", "fourth quarter 2008 earnings expectations" and "IBM" (the latter two expressed as their respective Semantic Maps) might be linked to seek information from multiple documents on the consensus projected fourth quarter earnings for IBM this year. Notice that there are two different "kinds" of concepts in use here. First, there are the concepts to be used to search through the documents (which are instantiated in Semantic Maps) and second there are concepts about what to do with that information. Thus "consensus" might mean "average all the data points you find in this document set". The information returned by the search may be of several kinds. Three important ones are where the information was found, what the data within that information (if any) was, and how strongly were the concepts expressed. Information location comes from the indexing process. When a concept has been found in an indexed document, its location is known (since that was determined as part of the indexing process). The data extraction may simply present the result (e.g., showing it within a document), or it may involve aggregating data from multiple sources in complex ways.

Conclusion

RAF has built a compiler for concept searches which is currently being used by Radiant Metrics to manage their hedge fund. Our approach does not apply full natural language processing because we have determined the complexity and maturity of this approach would simply be difficult to implement and therefore possibly generate too many false positives. Instead, our compiler uses statistical natural language processing which takes the structure of language, and translates it into bits and bytes of useful information. RAF has bridged the gap between words, parts of words, and concepts which provide users with the capability to search and capture the real meaning of the information they seek.

For more information

To learn more about our Semantic Search Technology, please contact our sales department @ (425) 867-0700 or visit www.raf.com.