for Government

A White Paper

For more information:
Herbert L. Roitblat, Ph.D.
Dolphinsearch, Inc.
474 E Main Street
Ventura, CA 93001
805 585-2102 x124

DolphinSearch for Government

The patented DolphinSearch neural network technology provides highly scalable, fast, accurate, and comprehensive access to most forms of text in any language. The system automatically reads text documents and understands the content in context. This technology provides core capabilities wherever text-based material has to be stored and retrieved, recognized, compared, or classified.

DolphinSearch learns the meanings of words from the documents it reads. No preplanning or knowledge engineering is required. You do not need to expend time and money building taxonomies, thesauri, or other ontological tools. Rather, these tools can emerge from the content of the documents itself, as understood by the brain-like processing system that is the foundation of DolphinSearch. DolphinSearch finds documents based on meanings, not just the words. Its powerful relevance ranking algorithms bring the most relevant documents to the top of the list where they can be quickly accessed and exploited for their actionable information.

DolphinSearch applications include:

- Intelligence analysis (open source and classified)
- Content-based selection and filtering
- Email monitoring and archiving
- Smart indexing of technical manuals
- Content-based message routing
- Self-service user support systems
- Declassification reviews
- Compliance monitoring
- Litigation support
- Document management and categorization
- Incident report data mining
- *Wherever text has to be stored and retrieved, recognized, compared, or categorized*

DolphinSearch is easily integrated into an agency's workflow. It finds information that other data mining tools miss and organizes it according to the useable relevance of the documents. In a test for an intelligence agency it was compared with a common word-search tool. DolphinSearch returned superior results. For example, DolphinSearch provided the answer to a question in each of the first eight documents that it returned. The other system did not answer the question until document 13.

Systems that rely on entity recognition are useful, but are limited to recognizing the relationship between named objects, such as people or organizations. DolphinSearch is not limited to this kind of relationship. It can find meaningful relationships among concepts and ideas as represented within the documents. For example, in one case it discovered the relationship between named co-conspirators. In another, it discovered the relationship between different drugs used for a similar purpose. In another it discovered the relationship between "budgets" and "quarterly operating results."

Systems that rely on thesauri recognize that there may be more than one way to express an idea. These systems are limited, though, by the amount of work it takes to build a thesaurus that has your organization's word usage. Is "support" a synonym for price fixing or is it a synonym for stanchion? A generic thesaurus can actually make the problem of information retrieval worse, because it can drag in irrelevant meanings. Specific thesauri tuned to your particular vocabulary require a great deal of effort to construct. Other limitations are described below.

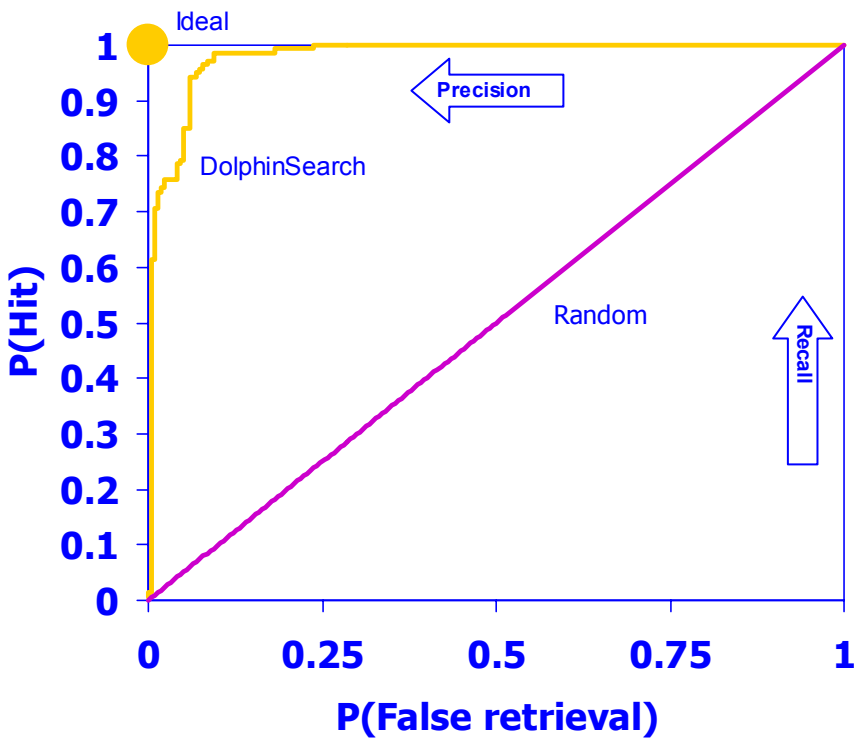
Systems that rely on categorization as the means for information management are limited by the brittleness of taxonomies. It is often difficult to say which category a particular document belongs in. People differ substantially in how they categorize a given document, depending on their needs and context. The same person can assign a given document to different categories at different times. We believe that categories are the outcome of an information management process, not the mechanism by which to achieve it. Categories help people to browse information, but they are not adequate as the only way to organize and retrieve it.

The effectiveness of DolphinSearch

DolphinSearch finds documents that other systems simply cannot find. DolphinSearch does not just count the number of times searched-for words occur; rather it uses proprietary *semantic profiles* to compare documents with one another or to compare documents with queries. A semantic profile is a mathematical representation of the meaning of a text object (a word, a sentence, a paragraph or a document).

The figure on the next page shows one way to think about the effectiveness of an information retrieval tool. It is customary to talk about precision and recall as measures of information retrieval effectiveness. An ideal system will have perfect recall, retrieving all of the relevant documents in the document collection, and perfect precision, returning only the relevant documents in the set. In the figure, the ideal system is located in the top left-hand corner.

The notions of precision and recall, however, are best applied when the system returns a precisely bounded set of documents. A system such as DolphinSearch, on the other hand, retrieves a fuzzy, but ranked list of documents. The top documents on the list are predicted to be most relevant to the search and the lower ranked documents are predicted to be less relevant. Precise values of precision and recall depend on where you place the cutoff—the rank at which you stop counting subsequent documents as relevant. Displaying the data this way highlights the tradeoff between precision and recall. One can increase the number of relevant documents retrieved, but at the expense of also retrieving more irrelevant documents. Traditional information retrieval systems are also

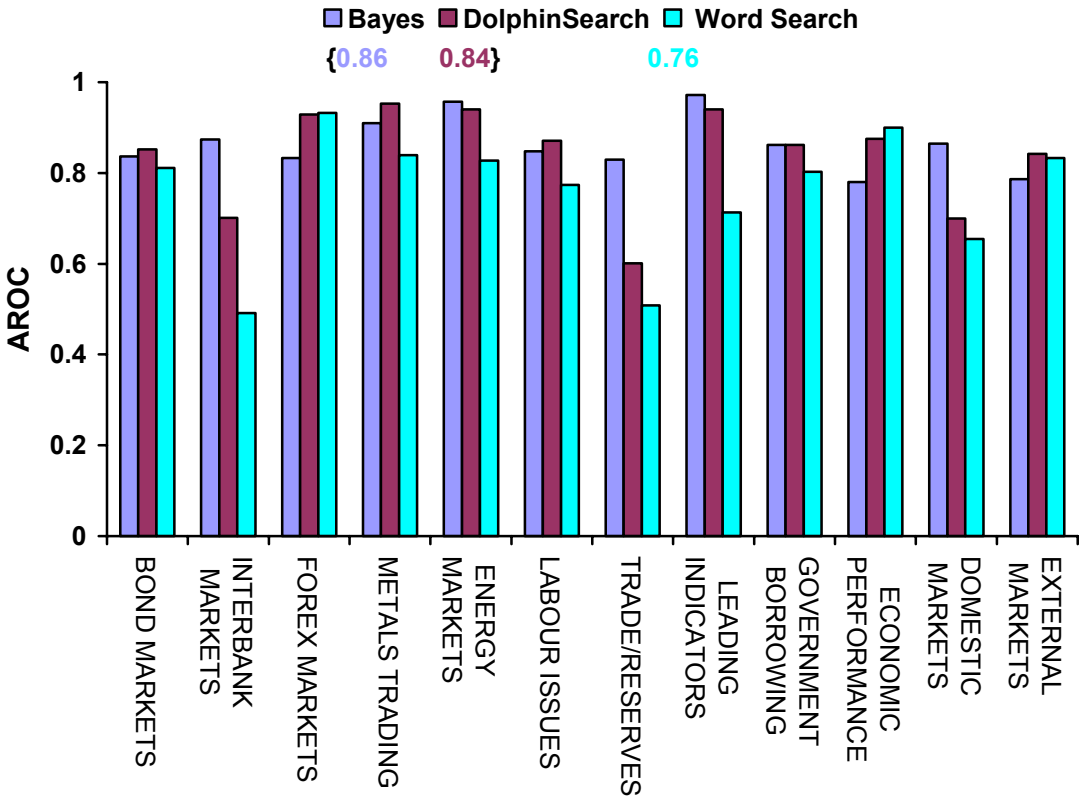


subject to this tradeoff, but it is often hidden within the design parameters of the system rather than presented to the user. Where a user will place the cutoff will depend on the value of the information being retrieved and on the costs of reviewing irrelevant documents. These costs and values may change as a function of individual queries, and so, we believe, should be made available to the users.

The data in this figure were derived from a small set of several hundred documents that were categorized for relevance by the agency that provided them. If one retrieved documents at random from the collection, then the performance of the system would fall somewhere on the diagonal line. A word search tool would produce a curve somewhere to the left of the diagonal, though the explicit comparison was not available for these data.

The curve represents the cumulative proportion of relevant and irrelevant documents retrieved at each rank in the list. DolphinSearch returned a few irrelevant documents mixed in with the relevant ones, but by document 192 on the list, DolphinSearch had retrieved all 140 relevant documents in the collection.

One can summarize the performance of this system by calculating the area under the curve for a specific system. This area is a summary of both precision and recall and explicitly recognizes the tradeoff between the two. A random system has an area of 0.5, a perfect system has an area of 1.0. DolphinSearch yielded an area of 0.97.



These results were obtained with a small set of documents. With larger sets it is much more difficult to get the ratings needed calculate precision and recall or to draw this curve. We expect that performance on large datasets will be a little lower than described here, but will still be better than those obtained with standard data mining or information retrieval tools.

The next figure shows the same kind of analysis with a set of commonly used test documents. The height of each bar represents the area under the ROC curve for the corresponding system and category. These were news articles selected from the Reuters Corpus Volume 1 (RCV1). The categories were assigned by Reuters coders using a combination of manual and automated (rule-based) coding procedures.

The data in this figure represent the performance of three different information management tools. The first is a Bayesian classifier trained on ten example documents from each category. A Bayesian classifier is a statistical system that computes the probability that a document belongs to a certain class given the words in that document. It requires someone to determine the categories that are to be classified and then to find example documents of each of those categories. If you want to add another category of documents after training is complete, you have to find examples of that new category and then train the entire system from scratch.

The second information management tool shown in the graph is DolphinSearch. DolphinSearch did not require any human-designed pretraining. DolphinSearch is equally capable of categorizing these documents into the predetermined categories as well as into any novel categories that one may need without any further effort.

The third tool was a word search engine. The same queries used for DolphinSearch, which were often the name of the category, were used for this word search engine.

DolphinSearch and the Bayesian classifier performed equally well (AROC: 0.86 v. 0.84) on categorizing the Reuters articles into their designated categories, despite the fact that no preplanning was required to use DolphinSearch. The word search engine performed significantly more poorly on these data. DolphinSearch, in other words, was as effective as a system that requires large amounts of up front work and superior to a system that used standard technology to find relevant documents.

Another DolphinSearch feature that makes the system easier to use than other systems is the ability to "explode" a document. Once a document has been retrieved, the system can jump to the most relevant section of that document. One does not have to read pages 1-99 of a document to find that the relevant information is on page 100.

In various tests, DolphinSearch has:

- Identified co-conspirators
- Recognized related chemicals and drugs
- Identified paraphrases of key terms
- Learned acronym expansions
- Improved relevance

Why word search is inadequate

Anyone who is familiar with searching on the internet knows what a standard search engine returns and what its limitations are. The poor performance of the word search tool described earlier is characteristic of the performance to be expected with this kind of system. Words have multiple meanings and an ordinary search engine has no way to resolve this ambiguity. The top 500 words of English have an average of 23 definitions each.

The following sentence is a paraphrase of an Adams saying. The numbers below each word are the number of definitions for that word.

This	form	of	government	seems	like	the	best	one	man	can	devise.
8	45	16	8	5	32	2	20	24	26	6	4

If you combine all possible senses of these words, there are 4,416,602,112,000 (4.4 trillion) possible interpretations. Competent English readers have no trouble understanding this sentence but computers do not ordinarily have any way to reduce the number of possible interpretations. For a human, each word disambiguates the other words in the sentence and lets the human come to some understanding of the sentence as a whole. Lacking such contextual information, standard search solutions must treat each word as an independent object with all of its attendant ambiguity. So, when you search using one of these words, an ordinary system returns documents with every possible meaning for each word, not just the one you had in mind. DolphinSearch employs a powerful theory of meaning to work toward mimicking the processes used by the brain and resolving the ambiguity.

Synonymy is another problem with using just word search. There are many ways of saying essentially the same thing. For example, there are more than 120 words that refer to thinking (assess, evaluate, interpret, contemplate, etc.). Other concepts are similarly variable. For example, there are said to be about two dozen expressions to describe the balls of fluff found under a bed, including "dust bunnies," "fairy flop," "dust kittens," "fluff bunnies." You buy liquor from a "liquor store" in New York, a "package store" in much of New England, and a "state store" in Pennsylvania and New Hampshire. People may speak obliquely about an issue. They may have special nicknames or acronyms. The head of a company might be the CEO, the Chief Executive, the President, the Chairman, the Director, or the Principal. He or she may be referred to by other names as well. A merger might be called a buyout, a joining, a merger, a consolidation, etc. The bottom line is that it is very difficult to anticipate how the author of a document may have discussed a matter.

People could use any number of different words to refer to the same idea, but they are generally poor at remembering exactly which specific word was used in a given context and equally poor at guessing what words another author may have used. For example, when looking for the smoking gun in an antitrust case, one of the following might be the actual sentence that the company president used, but how do you find it among many other innocent statements about Company X?

- "How much do we need to pay you to screw Company X? This is your lucky day."
- "What do you want to kill Company X? This could be the day your ship comes in."
- "Let's make a deal, maybe the best deal of your life. You smash those guys and we'll make it worth your while.

Two related problems could be called the Jabberwocky Effect and the Humpty Dumpty Syndrome. The Jabberwocky effect is the practice of using new words. People are very creative at making up new jargon to use in their discussions with one another. The Humpty Dumpty effect is the practice of using old words in new ways. Every organization has its jargon and acronyms that introduce new words and use old words in new ways. Understanding these new words and new uses and identifying the right words to search for presents a challenge to any information management process.

Word search alone would not be enough to find the document unless you were very sure that you could think up all of the ways that somebody could say something, which, outside of certain technical domains (perhaps), is an almost impossibility. You might be able to come up with a reasonably-sized list of key concepts for a case, but it is less credible to think that the names of these concepts will make adequate query terms for finding documents that exemplify those concepts. This speculation was supported by a study that found that attorneys were only about 20% effective at thinking up all of the different ways that the document authors could refer to issues in the case.

The case involved a San Francisco Bay Area Rapid Transit accident in which a computerized BART train failed to stop at the end of the line. There were about 350,000 pages in about 40,000 documents for the case (Blair and Maron, *Communications of the*

ACM, 28, 1985, 289-299). The attorneys worked with experienced paralegal search specialists to find all of the documents that were relevant to the pertinent issues. The attorneys estimated that they had found more than 75% of the relevant documents, but more detailed analysis found that the number was actually only about 20%. The authors of this study found that the different parties in the case used different words, depending on their role. The parties on the BART side of the case referred to "the unfortunate incident," but parties on the victim's side called it an "accident" or a "disaster." Other documents referred to the "event," "incident," "situation," "problem," or "difficulty." Proper names were often not mentioned. The limitation in this study was not the ability of the computer to find documents that met the attorneys' search criteria, but the inability of the attorneys and paralegals to anticipate all of the possible ways that people could refer to the issues in the case.

Concerning one issue, the attorneys in the case identified three terms that they thought would be adequate to retrieve relevant documents, Blair and Maron found 26 more. The original three words could not by themselves be used effectively to find relevant documents, because they retrieved too many irrelevant documents. Other search terms were needed to limit the range of documents that were returned, but this limitation came at the cost of missing documents that did not happen to have these additional terms. Coming up with the right combination of terms to yield relevant results and no irrelevant results is nearly impossible.

They found that the terms used to discuss one of the potentially faulty parts varied greatly depending on where in the country the document was written. Some people called it an "air truck," a "trap correction," "wire warp," or "Roman circle method." After 40 hours of following a "trail of linguistic creativity" and finding many more examples, they gave up trying to identify all of the different ways in which the document authors had identified this particular item. They did not run out of alternatives, they only ran out of time.

How DolphinSearch learns

Part of the DolphinSearch setup is to point it to the repository of documents that it is supposed to learn about and index. The documents are loaded onto a file server and DolphinSearch goes about reading each document and extracting the text. It then breaks up each document into paragraphs of text and transforms each paragraph into a mathematical form that can be used to train the neural network. As the philosopher Wittgenstein pointed out, the meaning of a word is determined by how it is used in the language. When an adult human learns a new word, it is generally learned in the context of other words. As we mature, the meaning of a word is given by the context in which it is currently being used and the history of contexts in which it has been used.

In the sentence, "The fortune teller examined the young man's palm," we do not have any problem understanding that the word "palm" refers here to the fellow's hand. In the sentence, "The tree surgeon examined the young man's palm," we similarly do not have any trouble knowing that the word "palm" refers here to a tree. We can know these two different meanings because the neural network in our brain has learned about how the word "palm" has been used in our experience and can see how the word "palm" is being

used in the context of the sentence. This is the kind of learning that makes DolphinSearch work. It is the kind of learning that has heretofore eluded computers.

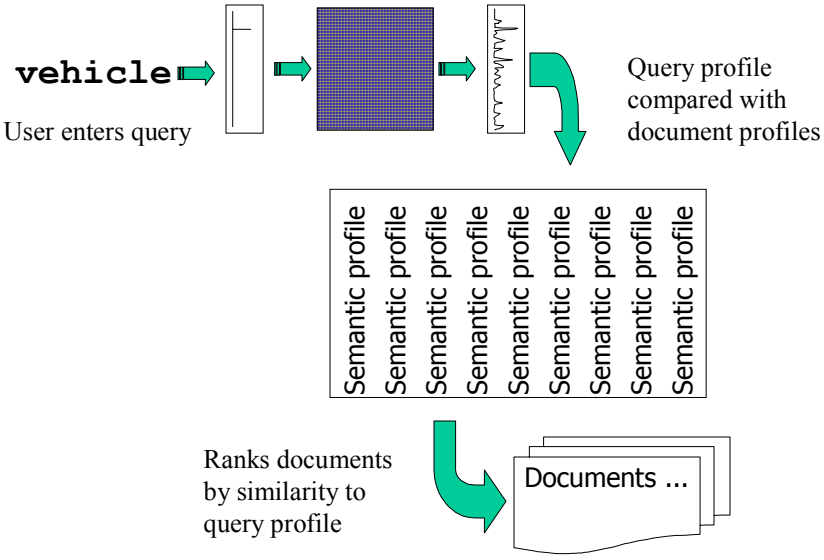
The DolphinSearch neural network forms strong connections among words that are strongly related to one another—words that occur together frequently and make up the same meaning-context. As it reads more documents, it learns more and more about how each word is used in the context of those documents. Spurious or adventitious relations are "washed out" leaving behind only the most important and most dominant meaning relations. As a result, in a collection of documents about bankruptcy filings, it will learn that bankruptcy is related to finance, court, judge, and receiver, among other things. This context is part of what DolphinSearch knows about the words and part of what controls how it responds to queries. Keep in mind, that no separate training set of documents is required. Under ordinary circumstances, DolphinSearch learns from the whole set of documents that it reads. If necessary, though, one can feed specific documents to the system as the training basis.

How it Works

DolphinSearch employs patented neural network technology to read the documents, extract their meaning, and make it all searchable. DolphinSearch forms semantic profiles for each word. A semantic profile is a mathematical representation of the meaning of the word, in relation to all of the other words that they system knows about. These semantic profiles capture meaning and reduce ambiguity. When you do a content search, DolphinSearch compares the semantic profiles of the query against the semantic profiles for each of the documents. It then ranks the resulting document list by the degree of similarity between the semantic profile for the document and the semantic profile for the query. The result is a true fuzzy semantic search.

Everything DolphinSearch knows about the meanings of words it learns from the documents that it reads. Acronyms, and words that are used in a unique way in the documents are learned as they are used, not confused with more generic meanings. For example, in one demo, we trained a DolphinSearch system on Japanese golf articles (articles in Japanese about playing golf). A reviewer entered two queries, one about the word in Japanese meaning "tired" and one about the Japanese word for "fatigue." These words gave different results, and he wondered why. On investigation, it turned out that "tired" was always used in these documents to talk about golfers being tired at the end of the day. "Fatigue," on the other hand, was always used to talk about metal fatigue and stress fractures. In the context of those documents, these two words were not synonyms. If one had built a thesaurus ahead of time making these two words synonyms, the system would have returned erroneous results.

When most systems talk about fuzzy searching, they are talking about tolerating spelling errors and word form variations. DolphinSearch, in contrast, employs fuzzy semantic searching. Fuzzy searches are an advancement over standard search mechanisms. In a standard or crisp search, a document either matches the criterion or it does not match the criterion. In a fuzzy search, the documents match to varying degrees. Some documents match very closely and are considered to be most relevant. Other documents match less closely and are considered to be less relevant to the query. DolphinSearch searches are fuzzy in this sense as well. They are semantic in the sense that we compare word



meanings rather than compare word occurrences. The semantic profile represents the meaning of the words in the context of the documents and the search uses the context of the words in each document as well as the searched-for words to determine the degree to which a given document is relevant.

As the figure shows, the system accepts a query and forms a semantic profile for it, by passing it through the neural network. It then compares this query semantic profile with the semantic profiles for each of the documents it has read. The result list is then ranked by the degree to which the document matches the semantic profile. The best, closest matching, documents are returned first in batches whose size is user selectable.

DolphinSearch allows users to perform Boolean searches as well as fuzzy semantic searches. In many cases, however, concept searches will be the most appropriate way to search for relevant information because the search results are ranked according to the meaning of the terms as they are represented by the documents rather than just by the occurrences of specific words. Use phrase searches when the words in the phrase mean something different than they do independently.

A rough guide to relevance ranking

DolphinSearch ordinarily ranks the documents returned in response to a query in relevance order. Relevance is the degree to which the semantic profile for the query matched the semantic profile for the document. Closer matches are considered to be more relevant. Practically, this ranking usually means that the first documents returned are those that contain the search terms in the context that DolphinSearch has learned to be most dominant for those terms. Next are usually documents that contain the search terms in different contexts from the one that has been learned. Finally come documents that contain the context without the specified search terms. DolphinSearch assumes that you

mean what you ask for, so it will do its best to return documents relevant to your query. Remember that all it knows about your query, though, is what you have typed in.

Analysis of the search results obtained from a document set about antitrust cases will help to illustrate the relevance ranking order. A substantial number of documents in this set concerned an antitrust action involving two companies that owned a large number of grain elevators (where grain is stored). A search for the word "elevator," turns up a number of documents that contain this word.

The first 15 documents are about grain elevators, the 16th and 17th documents mention elevators in the context of escalators. The 18th document does not mention elevators at all, but does mention the two firms whose merger was the subject of this litigation. The next seven documents are also about this same case, even though they do not contain the word elevator at all. A search for the word "elevator" found documents relevant to the case whether or not they contained that specific word. It also ranked those documents in a relevance order that was highly dependent on the context in which the search word was used. DolphinSearch delivered highly relevant results, based on their meaning, rather than based on how many times the document happened to contain the searched-for terms.

Conclusion

DolphinSearch technology drives a system that is based on machine understanding of the text that it reads. The DolphinSearch solution is highly scalable and cost effective.

DolphinSearch is easily integrated into the workflow of an organization. Its web-based architecture makes it accessible over both secure networks or even over the internet. It learns from the documents that it reads, so there is never a need for complex knowledge engineering to get a usable and effective system. It works in any language.

DolphinSearch provides a powerful information retrieval and data mining tool that gives access to a new standard of high quality information. It has achieved commercial success by making it easier and faster for clients to find the information that they need in a form that they can use for action. Relative to traditional word search, manual tagging, or other technologies, DolphinSearch delivers far superior information to its users when they need it, without the need to preplan, prestructure, or precategorize information. Its operation mimics that of the brain, rather than forcing users to mimic the operation of traditional computers.