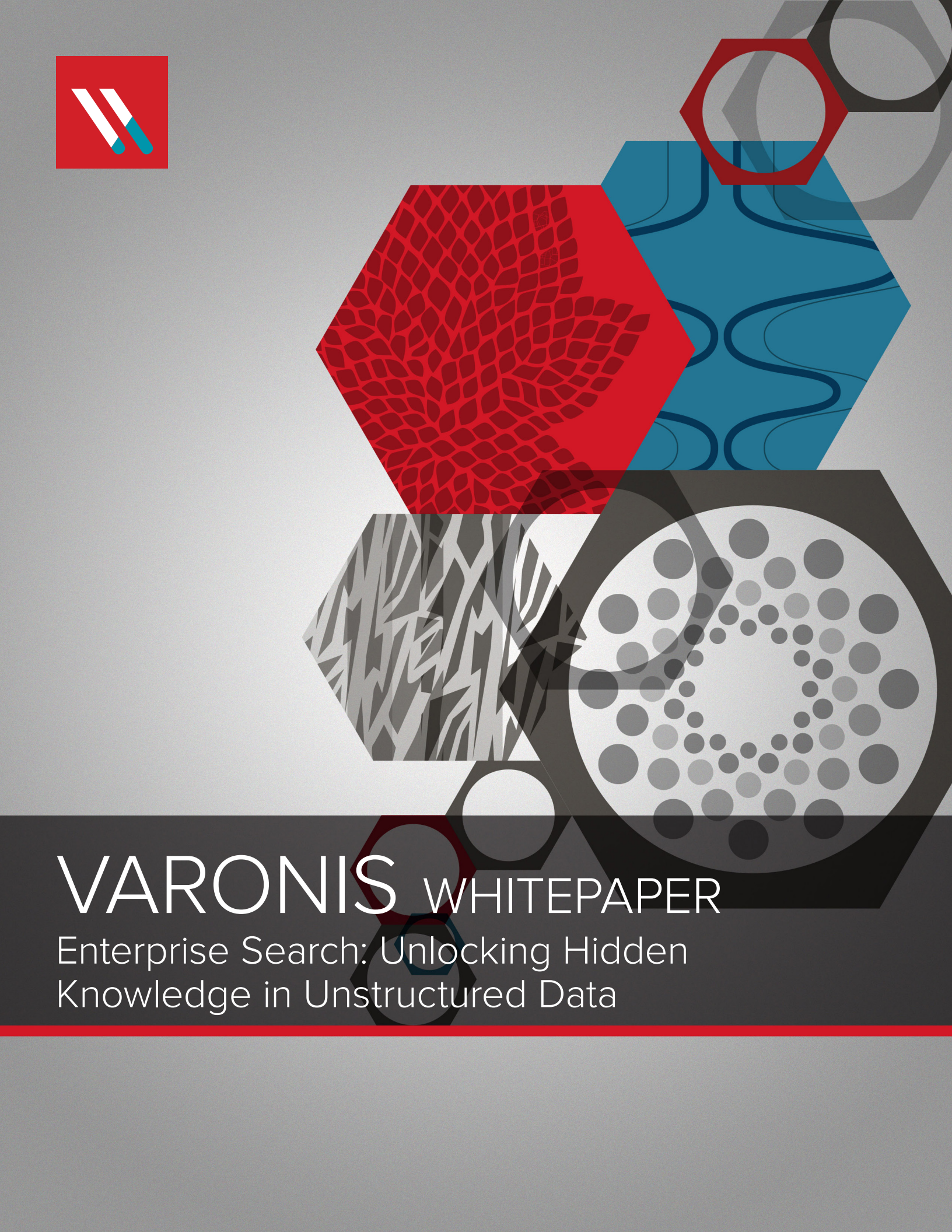


VARONIS WHITEPAPER

Enterprise Search: Unlocking Hidden Knowledge in Unstructured Data

CONTENTS

ENTERPRISE SEARCH: UNLOCKING HIDDEN KNOWLEDGE IN UNSTRUCTURED DATA

OVERVIEW

The major search engine players–Google, Microsoft, Yahoo–have revealed little about their search algorithms. That's understandable: it's their secret sauce. However, from the few statements they have made publically, we have some understanding of how the underlying algorithms work in finding most relevant webpages.[i]

At a high-level, they are effectively scanning and indexing all the world's websites–there are now approximately 1 billion sites[ii] . You can think of the search engine as simply taking the query words we enter in the search box and then matching against a humongous dictionary-style index. Google was certainly far better than others at *ordering* the results from the matches using its now famous PageRank algorithm, which cleverly exploited web-based metadata[iii].

But Internet searches have begun to move to the next stage. Search engine companies have long recognized that people are not looking for just a list of sites in the search results, but actual answers to questions. This is being called **semantic search**–viewing the query as a question that requires results that look more like answers.

So when entering "empire state building height", you're really looking for the height of this iconic landmark and a number. And Google and Microsoft can readily answer this question–1,454 feet, by the way.

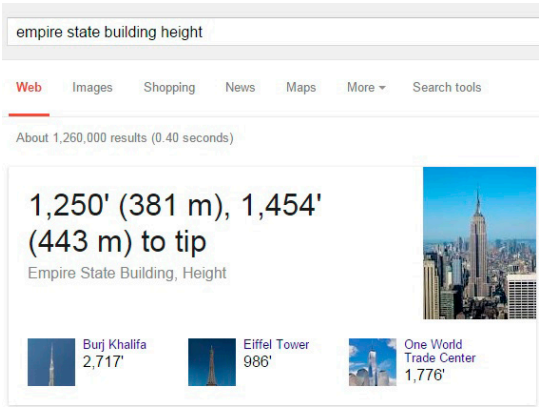
Enterprise search is also now ready for a similar revolution. Just as in the Internet world, employees want answers to questions, "who are the sales reps in Iowa?", in addition to finding basic answers provided by first-generation enterprise search.

And enterprise search can leverage the same ideas used by Google and others by also treating the query words as a question and their file systems, intranets, and even email as a source of knowledge.

As it does in the Internet world, metadata plays an important role in fine tuning queries and interpreting unstructured file data as useful information. In this paper, we'll explore how some of these same ideas can make enterprise search better and smarter.

THE NEW SEARCH ENGINE: BRILLIANT SUGGESTIONS, BETTER ANSWERS

Search engine gurus who spend their days thinking about how to better deliver relevant information have a big picture view of the search process. It goes something like the following: On the Internet, we enter search terms to do one of three things: find a website by name, find information or even answer a question, or engage in a transaction–sometimes referred to as 'do, know, go'.[iv]



The second category, informational, makes up most of the queries received by Internet search engines and is a subject of continuing research. Who hasn't used a search engine to get both basic information (names, addresses, dates) and, more recently, deeper information about, say, the best Italian restaurant in your neighborhood, or the height of a famous NYC building?

Search engines try to discern the intention of the person entering the keywords –what he or she is really asking for–so it can be slotted into one of the above categories. Search engines then typically do further query analysis to get a more granular classification.

For an informational query, the engine might decide to just point the person to an appropriate web article or deliver a specific answer (the height of the Empire State building.)

How does the search engine make this decision? One of the techniques that the search engine vendors have come up with to help untangle user intent are the helpful autosuggestions that drop down from the search box.

AUTOSUGGESTIONS: FINDING THE RIGHT SEARCH WORDS

We've all experienced these automatically generated keyword tips–who hasn't relied on the autosuggestion to correct misspelled names, correct our own malformed searches, and fill in additional information?

While we don't know the exact formula that the search engine companies use to derive autosuggestions, researchers and others have come up with a good list of what they're likely using[v]:

- search volume of similar keyword queries
- location
- freshness of the query
- and web mentions

It's also a good time to point out that these factors are a form of metadata and that metadata has similar powers in an enterprise search environment–see our Next Generation Enterprise Search to learn more how file activity metadata can be used rank results from enterprise search.

Let's now go through some of these factors so we can get a better feel for how they can help find answers.

The frequency with which the same search keywords are used by everyone else is quite helpful in tuning the search words. Makes sense! If many people are entering similar searches, then you have a reliable window into what people are really interested in. Timeliness, or "freshness," also helps–what people are searching for at the current moment.

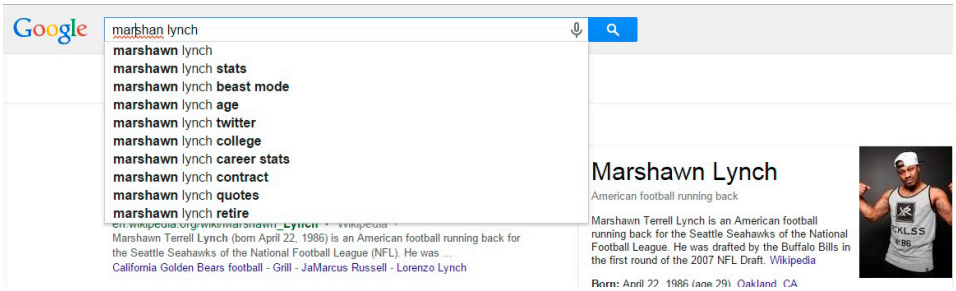
For example, on Sunday February 1, or as its referred to here in the US, Super Bowl Day, Google responded with the following auto suggestions when the keywords "Super Bowl" was entered: Super Bowl start time, Super Bowl kickoff time, Super Bowl 2015. You can see what the analysis algorithms are doing here: trying to help you decide whether you want a specific answer to a common question or broad information.

Search engines also further refine keyword popularity based on location–likely at the state level. In the east, when you enter ski areas, Google will suggest "ski areas Vermont." Out west, the suggestions become "ski areas Utah".

In the above cases, Google and other players rely on the "wisdom of the crowd" what everyone else is entering provides a good guide to understand what a particular search engine is being used for.

But there's also wisdom in the data itself.

When Google notes that a keyword matches (or is close to matching) very common word usage in the web pages, it will suggest that option as well. Type in “marshan lynch”and Google will suggest “marshawn lynch” the Seattle running back–there’s just more content on those keywords–so it assumes you really want to learn about those.



HELP FROM SCHEMAS FOR MORE BRILLIANT SUGGESTIONS AND RESULTS

As anyone who’s been following Google and other search engine players in the last year, you may have noticed that the suggestions and the searches have gotten better.

In Google, for example, enter “charles dickens born,” and the autosuggestion is “charles dickens born date”, and Google provides the actual answer!

How does Google know all this?

We have a knowledge map in our minds. We know that A Tale of Two Cities is a novel, which is a type of book, and books are associated with authors. Charles Dickens is an author, an author is a person, and people have dates associated with them: data of birth, date entered college, etc.

Google and the other search engine players have their own digital versions of the knowledge map[vi]. Computer scientists refer to it as a **semantic schema**, which provides the skeletal structure for organizing information. Webmasters dip into the schemas to direct how they markup their HTML. Semantically marked up HTML tells Google and other search engines the meaning underlying the content. For example, the “rich snippets” that Google provides is one example of how it takes advantage of marked-up content.[vii]

What else is Google doing? It has created a separate searchable knowledge base, which it calls the Knowledge Graph[viii]. It’s in this knowledge base that the answer to the age of Dickens (and lots of other famous people) can be found. Knowledge graphs help Google create brilliant query suggestions and even more brilliant answers.

Can enterprise search perform the same kind of knowledge magic?

ENTERPRISE SEARCH: METADATA AND KNOWLEDGE

The answer is yes! However, some of the ideas we've just discussed on embedded knowledge may not be as significant (yet) in an enterprise environment as they are in the Internet world. After all, employees are generally looking to find files or other content–the enterprise equivalent of finding the name of website.

Still, our expectations are always increasing with consumer technology now pointing the way. It won't be long before we'll demand the full power of search–just as we want our consumer devices available at work with BYOD policies.

The key is that enterprise search software has to have access to file metadata, which plays a critically important role in making the results even smarter.

Remember how location information of Google searchers was used to help complete queries on skiing? If you live in New Jersey, Google might complete "ski resorts" with "ski resorts Vermont".

What would the parallel be to location in an enterprise environment?

The answer: clusters of users and groups based on groups and departments that are maintained in Active Directory, actual usage, past searches, and content. I'd certainly want my autosuggestions shaped by what others on my team or department are entering, using, or searching for.

If the marketing team has been searching frequently for "product roadmap spreadsheet", "product roadmap project Atlas", and other related terms, then a cruder search, like "product roadmap," could be expanded based on the crowd wisdom contained in other, similar marketing searches.

In an enterprise environment, it would also be helpful to know the file access activities of all employees and organize those with similar patterns, regardless of which department they belong to, into the same virtual group. Of course, this would require deep knowledge of user file activity patterns–available if you have the metadata.

These groupings could then be used to adjust the results!

So for example, suppose you belong to the marketing group, but spend a lot of time accessing files under the sales and competitive research folders. With autosuggestions guided by analysis of activity metadata, results could be weighted more heavily on the keyword that are popular with the sales group. Neat idea.

But what about pulling knowledge out of the data itself? No one's expecting employees to markup their data to enable semantic searching–as webmasters do for web content. However, lots of content does have an implicit structure!

Consider all the information that can be gathered from lists and tables in Word documents and PowerPoint presentations. Or from lists or bullet points that are organized under topic sentences.

Sales team	Region
Bob	Midwest
Frank	East Coast
Rich	Canada

With a backend that can parse, classify, and interpret all this hidden structured information contained within employee unstructured content, there's great potential to match queries to real answers.

Suppose an enterprise search engine while indexing and analyzing the content comes across a table like one shown above. If an employee enters "sales team" as query words, the enterprise search can suggest "sales team bob" or 'sales team frank" and then bring up highly relevant results.

If it understands–thanks to a semantic schema–that the query words refer to employees, it might dip into Active Directory and pull out relevant contact information. And if the enterprise search had access to a knowledge base of general facts, it could supplement the answer with geographic and other contextual information about team members.

CONCLUSION: THE FUTURE OF ENTERPRISE SEARCH

Prior to the Internet revolution, most companies' experience with finding and searching for important content was through closed-off document management systems. You had to explicitly import files into the 'DMS', and then search for file names and content using their rudimentary functions.

The new generation of enterprise search goes well beyond these capabilities. The search model is similar to the one blazed by Google and other search providers. The file system mirrors web pages, and employees are both general consumers of content as well as contributors.

To complete these connections enterprise search can use an analogue to web link metadata, which is the extra information for telling the algorithms what content is more important than others. In enterprise search, this metadata is found in the file system, Active Directory and in employee file usage.

Metadata about employee file access patterns can also help shape search queries through better autosuggestions. And metadata plays an important role in pulling out the knowledge embedded in file system content. Google and Bing are already using similar web user metadata to improve their search functions.

We believe it won't be long before companies and employees demand the same capabilities for their in-house content search capabilities that they already experience with their Internet search engine providers.

[i] *https://support.google.com/websearch/answer/106230?hl=en*

[ii] *http://www.internetlivestats.com/total-number-of-websites/*

[iii] *https://en.wikipedia.org/wiki/PageRank*

[iv] *http://searchenginewatch.com/sew/how-to/2235624/do-know-go-how-to-create-content-at-each-stage-of-the-buying-cycle*

[v] *http://moz.com/blog/how-googles-search-suggest-instant-works-whiteboard-friday*

[vi] *http://schema.org/*

[vii] *https://developers.google.com/structured-data/rich-snippets/recipes*

[viii] *http://www.google.com/insidesearch/features/search/knowledge.html*

ABOUT VARONIS

Varonis is the leading provider of software solutions for unstructured, human-generated enterprise data. Varonis provides an innovative software platform that allows enterprises to map, analyze, manage and migrate their unstructured data. Varonis specializes in human-generated data, a type of unstructured data that includes an enterprise's spreadsheets, word processing documents, presentations, audio files, video files, emails, text messages and any other data created by employees. This data often contains an enterprise's financial information, product plans, strategic initiatives, intellectual property and numerous other forms of vital information. IT and business personnel deploy Varonis software for a variety of use cases, including data governance, data security, archiving, file synchronization, enhanced mobile data accessibility and information collaboration.

Free 30-day assessment:

WITHIN HOURS OF INSTALLATION

You can instantly conduct a permissions audit: File and folder access permissions and how those map to specific users and groups. You can even generate reports.

WITHIN A DAY OF INSTALLATION

Varonis DatAdvantage will begin to show you which users are accessing the data, and how.

WITHIN 3 WEEKS OF INSTALLATION

Varonis DatAdvantage will actually make highly reliable recommendations about how to limit access to files and folders to just those users who need it for their jobs.

START YOUR FREE TRIAL