Fuzzy Name Search for Due Diligence & KYC

Why Vital4 chose Rosette for business-critical search of people and companies in its due diligence platform

Executive Summary

Customer or money launderer? New hire or felon? Whether it's for financial compliance or pre-employment screening, due diligence is all about finding names and matching names against watch lists, negative news reports, SEC filings, and more. Vital4 is a standout data aggregator; providing and updating thousands of data sources 24x7 for due diligence vendors and corporate departments. Search—specifically for people and companies—is more than mission critical; it's the means of making all that data usable. Unsatisfied with their existing search, Vital4 went looking for better fuzzy name search and a better way to scour news about people and companies. They found the search experience their customers were demanding by upgrading their system with the advanced natural language processing within Rosette.

"The search we left behind was just not the kind of experience our clients wanted. It came down to what our clients were asking for, fuzzy name matching."

Search—specifically for people and companies—is more than mission critical; it's the means of making all that data usable.

WHY "REGULAR" SEARCH ISN'T BUILT FOR "FUZZY NAME SEARCH"

Most search engines use a scoring system called TF/IDF (frequency–inverse document frequency) to find the most relevant result. This method assumes that terms which appear more frequently in a document are less important. However, this standard search method for documents is not effective for finding names. Common words such as "the" are considered insignificant within search. However with names, the very common name "John" is just as vital as the unusual "Dweezil."

Fuzzy matching for names matching is a very particular matching problem. Misspellings in documents like "htere" are easily "corrected" by search engines, and matter less, because they are such a tiny part of the whole, but when it comes to names, is "Cyndy" really a misspelling of "Cindy" or a completely different person? That makes a big difference.

In brief, name search is a very particular search problem that requires a very specialized solution.

Vital4: Tech-Driven Due Diligence

Vital4 fills a key role in due diligence, supplying aggregated data essential to platform vendors or compliance departments who rely on the data to provide screening for areas such as pre-employment, KYC for financial institutions, and M&A.

The data Vital4 provides is not just from watch lists but also negative news gleaned from news articles around the world in many languages, and information about politically exposed persons (PEPs). Their strength is in aggregating thousands of sources more than their competitors and maintaining full FCRA (Fair Credit Reporting Act) compliance for U.S. customers, meaning that as soon as a person is off a watch list their name must be removed from their databases within 24 hours. Ingesting the data in whatever format it arrives and then making it accessible is a huge job.

The Problem

For Vital4 clients, all searches center on people and businesses. However, traditional keyword-based search did not provide the accuracy that Vital4 sought to provide. Limitations of their previous search solution included:

1. Lack of fuzzy name matching
2. Poor relevancy in searching for people and businesses in articles

First their previous search solution lacked true fuzzy name matching capable of matching names that were very likely "the same" but appeared different due to typos (Fuhrmann vs. Fuhrman), phonetic errors (Hawkinberry vs. Hockenbury), nicknames, initials, truncations (Chas. vs. Charles), misordered name components (John Henry or Henry John?), or were simply written in two different languages ("Ichiro Suzuki "and "鈴木一郎").

Fuzzy search on documents is in some ways simple; misspellings like “teh” are a minuscule part of each document and don’t affect a search for “potato pancake recipes” or require much intelligence on the part of the search algorithm. Names are much more difficult because the difference of one character can mean you have two different people or are missing a match. Is it “John Chu” the “Jon Chew” you were looking for?

Second, in searching news articles, a search on “John Smith” might hit an article because it contained, “Smith explained that John and his team were responsible for the damage…” The article would be returned as a hit only because of the proximity of “John” and “Smith,” but to a human, it clearly doesn’t qualify as an article mentioning a “John Smith.”

Manually having human editors tag every article with the people and companies mentioned would solve the problem, but is impractical when you have thousands of sources being refreshed daily.

> *Is it “John Chu” the “Jon Chew” you were looking for? If search were throwing snowballs, finding documents would be hitting trees, and finding names would be like hitting the tip of a small branch.*

The Solution

The company started by looking for fuzzy name matching and found Rosette, which offers two complementary capabilities to enhance their search:

1. Intelligent fuzzy name matching
2. AI-powered tagging of people and businesses in articles

Since Rosette plugs easily into Solr or Elasticsearch—both of which offer flexible search indices—it was an easy decision to migrate to Elasticsearch from their previous, rigid search system.

INTELLIGENT FUZZY NAME MATCHING

Vital4 clients expect the effortless experience they get from major search engines that can find relevant documents even with an imprecise query. Rosette offers Vital4 the ability to deliver highly relevant search results on names, that match the expectations of their users.

The greater precision of Rosette made it stand out. While many fuzzy name matchers rely on generating thousands of name variations in hopes of making a match, Rosette offers a sophisticated hybrid approach, using both phonetic matching and a statistical engine with a wide array of algorithms that contain knowledge about how names are written and pronounced in 15+ languages, and how they are transliterated across scripts.

For every name search of people or organizations, Rosette considers over a dozen variations including nicknames, abbreviations, truncations, and initials. Also, by flexibly comparing all the tokens between two names, Rosette overcomes any issues of misordered or missing name components.

13 WAYS ROSETTE MATCHES NAMES

Phonetic similarity
Jesus ↔ Heyzeus ↔ Haezoos

Transliteration spelling differences
Abdul Rasheed ↔ Abd al-Rashid

Nicknames
William ↔ Will ↔ Bill ↔ Billy

Missing spaces or hyphens
MaryEllen ↔ Mary Ellen ↔ Mary-Ellen

Titles and honorifics
Dr. ↔ Mr. ↔ Ph.D.

Gender
Jon Smith ↔ John Smith (but not Joan Smith)

Out-of-order components
Diaz, Carlos Alfonzo ↔ Carlos Alfonzo Diaz

Missing components
Phillip Charles Carr ↔ Phillip Carr

Initials
J. E. Smith ↔ James Earl Smith

Split inconsistently across database fields
Dick · Van Dyke ↔ Dick Van · Dyke

Same name in multiple scripts
Mao Zedong ↔ Мао Цзэдун ↔ 毛泽东 ↔ 毛澤東

Semantically similar names (cross-script)
Nippon Telegraph and Telephone Corporation ↔ 日本電信電話株

Semantically similar names
Eagle Pharmaceuticals, Inc. ↔ Eagle Drugs, Co.

Truncated components
McDonalds ↔ McDonald ↔ McD

Most importantly, every match is delivered with a match score to empower the user to decide what is "close enough" to be considered a match. That same scoring and weighting system integrates with the search engine's relevancy scoring algorithm to produce a single list of relevancy-ranked results.

"For name matching, we didn't see any other offerings that fit our needs as well as Rosette," Patrick Deeb, former CTO of Vital4 said.

AI-POWERED TAGGING

"Once we moved to a more flexible search index solution, it opened up the ability for us to make better use of the Rosette tools available to us," Deeb said. "We were able to start running all of our articles through entity extraction, and create an array of those entities to be searched on, rather than searching on the entire text of all the articles. This immediately added quality to the results."

An AI tagger is tireless and intelligent, leveraging the context to know when "Christian" refers to the designer ("Christian Dior") or the religion.

> *"We compared [Rosette] to some other text analytics and it seemed like the best. The search we left behind was just not the kind of experience our clients wanted. It came down to what our clients were asking for, fuzzy name matching."*
>
> —Patrick Deeb, former CTO of Vital4

Vital4 uses Rosette entity extraction which natively processes text in 20+ languages to automatically identify entities (people, places, locations, etc.) in each document, so that a search for "Paris Hilton" (person) does not hit a document about the Hilton hotel in Paris (place). The scope of due diligence is such that Vital4 has to consider news articles in every language for its database, and so it makes full use of Rosette's multilingual features.

Vital4 looked at other entity extraction solutions, but the choice of Rosette came down to its quality results. For icing on the cake, BasisTech offers a startup program which makes text analytics available to high-impact, early-stage companies.

"We compared [Rosette] to some other text analytics and it seemed like the best," Deeb said. "The search we left behind was just not the kind of experience our clients wanted. It came down to what our clients were asking for, fuzzy name matching."