BasisTech

Red Flag Group

More Relevant Search for Due Diligence

Red Flag Group saves time for clients by implementing Rosette intelligent name search

The Red Flag Group's global due diligence services cover 194 countries and 54 languages. Its full menu of services offers everything the compliance officer of a Fortune 1000 company needs: watchlist and adverse media screening; vendor reports; supplier onboarding risk scores; and compliance training. Red Flag Group puts together the technology and human expertise to predict, identify, monitor, and manage every risk to a client's integrity.

Red Flag Group's IntegraWatch platform allows compliance officers to screen the names of people and organizations against international sanctions and watch lists, a politically exposed persons database, a state-owned entity database, and adverse media from every country in the world. Their data is summarized and categorized into risk areas by human analysts who speak the local language and understand the lay of the land in a given country.

The Challenges

"The biggest frustration of our customers is sifting through all the irrelevant results [false positives]," Paul Johnson, Product Director of IntegraWatch at Red Flag Group, said. "The old system was a Soundex [phonetic search] that brought back a lot of false positives that seem to have no relationship to the desired record."

Getting too many irrelevant search results was expensive and laborious for clients to review. A given client might be monitoring hundreds or thousands of names using our ongoing monitoring capability, which rescreened search terms every 24 hours. This situation often resulted in new sets of "hits" or notifications that required remediation by the client.

> ***"The biggest frustration of our customers is sifting through all the irrelevant results [false positives]."***
>
> *— Paul Johnson, Product Director of IntegraWatch at Red Flag Group*

"Clients want to be notified when 'Acme Brick' in Russia was in the news for presenting some sort of client risk, but not for 'Acme Bricklaying' in France or 'Ace Brick' in Russia," Johnson said. "Ideally, we would only return 100% matches. That'd be the perfect system. The biggest dilemma is clients can't *not* look at all the results."

On the other hand, a missing match (false negative) in search results could be more painful than wasted investigations. A company doing business with a bad actor could damage its reputation or risk high financial penalties from regulators.

A client might search for thousands of names at once, so there isn't time to write name variations to catch cases where a name might appear with a slightly different spelling in one database versus another. Or worse, the database might list the name of a person or company, but write it in a different language or script (Nippon Telegraph and Telephone versus 日本電信電話).

"Multilingual searches were a main target we were trying to solve," Joseph Mantuano, Red Flag Group's Senior Developer, said. "For each name we would have had to collect every variation and every translation, but that's just too much work."

Thus, Red Flag Group started to look for a name-matching solution to address its pain points:

1. Too many false positives
2. Lack of fuzzy name matching requiring clients to input names and their variations
3. Lack of cross-lingual name matching.

Red Flag Group was using Elasticsearch for searches, and its default algorithm using TF/IDF (frequency–inverse document frequency) was not effective in finding the most relevant results.[1] Red Flag Group

[1]This method gives greater weight to terms that appear more frequently in a document. 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 common name "John" is just as vital as the unusual "Dweezil."

had a specialized thesaurus that claimed to do international name matching, but it was still only doing a straight comparison of strings (Johnson ↔ Johnson, Jonson ↔ Jonson).

> *"Multilingual searches were a main target we were trying to solve. For each name we would have had to collect every variation and every translation, but that's just too much work."*
>
> *— Joseph Mantuano, Senior Developer, Red Flag Group*

"Our older system performed very poorly on non-Latin languages such as Chinese, Arabic, and Japanese," Tim Hawkins, Red Flag Group's CTO, said. "The issue is some Chinese companies lead with their English name and sometimes the Chinese name, so there is a lot of inconsistency in how names are represented."

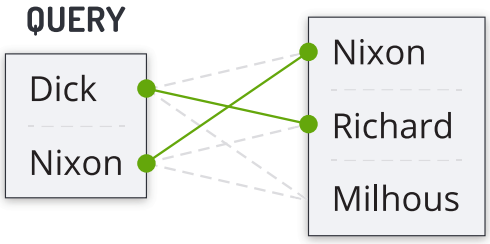
Furthermore, the name match score from Elasticsearch would change from time to time, depending on the size of the search index. Customers felt unsatisfied that a match might be scored 95% one day, but 90% the next day, after rerunning the same search.

Besides making clients happy, a better name search would give Red Flag Group a critical competitive edge in its industry.

The Solution

Hawkins and his team considered a larger-scale solution that offers name matching, but it was built for text mining. That was not an easy drop-in solution for their search optimization problem. While looking at the text indexer in MongoDB, Hawkins came across Rosette, which is used by MongoDB to expand language support of its search capability.

The Rosette name-matching features come as a convenient Elasticsearch plugin to add fuzzy name search across 18 languages (covering 12 scripts) and 13 variation types—including cross-lingual search. Essentially, at indexing time, Rosette encodes fuzzy matching lookup keys for each name to store in a hidden index field. Queried names are similarly encoded and enable fuzzy matching by matching the keys and calculating the similarity using machine-learning models. Further, within Rosette, names are broken up into tokens that are compared to determine the best match alignment.



Rosette compares names on a token-by-token basis, thus overcoming issues of misordered name components or names in the wrong database field.

The relative weight given each matching or "almost matching" token depends on language- and name-specific characteristics, such as the rarity of a name, same/different gender, and missing tokens. Note how the unusual name "Dweezil" carries more weight in Rosette's statistical HMM algorithm in calculating the match score for "Dweezil Jones" versus "Dweezil Johnson." By contrast, the more common name "John" adds less weight to the matching of "John Jones" versus "John Johnson."

WEIGHTING MODEL			
Finding a match for a common given name like 'john' is not as significant as finding a match for a rare given name like 'dweezil'. Although in isolation the token pairs have the same match score, in the context of the full name, these scores get weighted differently when the rarity of the tokens changes.			
*The match score for '**John Jones**' vs. '**John Johnson**' is **0.7902***			
'john'	'john'	MATCH	1.0000
'jones'	'johnson'	HMM_MATCH	0.4396
*The match score for '**Dweezil Jones**' vs. '**Dweezil Johnson**' is **0.9156***			
'dweezil'	'dweezil'	MATCH	1.0000
'jones'	'johnson'	HMM_MATCH	0.4396

For matching organizational names in some languages, Rosette also compares the semantic similarity of words in the name using text embeddings—one of the most powerful results of current deep learning research. It allows Rosette to match entities based on words with similar meanings, rather than only phonetics. For example, a search for "Eagle Drugs, Inc." of the NASDAQ database will fuzzy match "Eagle Pharmaceuticals, Inc." because "drugs" and "pharmaceuticals" are close in meaning.

Rosette hides all these complexities from Elasticsearch, simply returning a list of results by relevance with match scores calculated.

> *For matching organizational names in some languages, Rosette also compares the semantic similarity of words... [thus] a search for "Eagle Drugs, Inc."...will fuzzy match "Eagle Pharmaceuticals, Inc."*

As each user has different data and goals, Rosette offers several configuration options, such as balancing the recall (comprehensiveness of the search) and performance (speed) trade-off, or setting the minimum match threshold (e.g., 80% match) for a name to be returned as "a match."

The vast majority of queries from Red Flag Group clients are in English. But because Red Flag Group covers news sources in around 54 languages, name matches in multilingual content also need to be found. Some of the most important languages for them are Chinese, Russian, and Arabic. Rosette's support of language and cross-language name matching solves this problem.

Why did Red Flag Group choose Rosette in the end?

"The killer feature for us in Rosette was the ability to find Chinese names from English searches and vice versa," Hawkins said. "As customers do more business with China, there is more and more need to do searches on Chinese company names."

With the intelligent name matching of Rosette now inside IntegraWatch, Red Flag Group is gaining a significant competitive edge in customer retention and acquisition. Its customers feel confident that IntegraWatch has them covered, wherever in the world they do business.

ROSETTE FUZZY NAME MATCHING LANGUAGE AND SCRIPT SUPPORT

LANGUAGES	SCRIPTS
English, French, German, Hungarian, Italian, Portuguese, Spanish	Latin
Arabic, Pashto, Persian, Urdu	Arabic
Russian	Cyrillic
Chinese	Hanzi
Japanese	Hiragana, Katakana, Kanji
Korean	Hangul, Hanja
Greek	Greek
Thai	Thai

® 2022 Basis Technology Corporation. "Basis Technology" and "Rosette" are registered trademarks of Basis Technology Corporation.