rowanalytics

PI: **Serdar Uckun**

Key members:

**Steve Gardner,
Marcin Pawlowski**

The Problem

Search engines typically use keyword-based search systems with Boolean search. These engines do not scale well, rely on accurate and updated curation (tags) to be useful, and require expertise in query formulation. In life sciences, one study suggests that 95% of PubMed users fail to use search effectively (Illhoi & Mosa, 2015). To keep up with dynamically-changing corpora of information, we needed a smarter way to ~~search~~ **find** the content we seek.

Spot.my

Spot.my is a is a deep semantic search and matching platform that is much more useful for searching complex, nuanced data such as scientific literature, legal documents, or real-time news feeds. Spot.my and its underlying technology is a result of 20+ years of extensive research in semantic search and discovery. Currently, spot.my technology is used to serve two different domains:

- **Life science** (beta tool available at lifescience.spot.my)
- **News** (beta testing ended recently at news.spot.my)

spot.my
NEWS

spot.my
LIFE SCIENCE

• >5,000 RSS feeds	• based on PubMed® journal citations and abstracts for biomedical literature
• each feed updated every minute	
• full index updated every **40 mins**	• updated daily
• 1.6M indexed records	• 27M indexed records

The semantic indexing and search technology inside spot.my addresses several shortcomings of keyword-based search tools:

- Finds all documents of interest even if you don't know the right keywords
- Search using paragraphs of text or even documents instead of Boolean strings
- Semantic indexing engine is language-agnostic and it could simultaneously serve corpora in multiple languages. One could search using a paragraph in English and find relevant articles in a different language!

Key features:

- Domain independent – works with any source: structured databases, news channels, RSS Feeds, local document repositories
- Data type independent -- e.g., text documents, images, chemical components
- No training or manual ontology curation required!
- Horizontally scalable -- almost lineary scalable on multiple CPU and GPU systems
- Master index may be updated in near real-time with no impact on performance

Capabilities Sought

- Intelligence domain expertise
- System integration capabilities
- Visualization

Performance

We compared the retrieval performance of spot.my in the life science domain against two flagship products: Google Scholar and National Library of Medicine's PubMed ®.

1 **Spot.my** has better accuracy and precision Than Google Scholar and PubMed
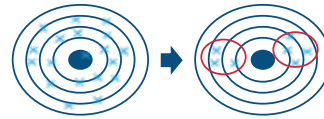
The first query was "*mechanism of disease*" and "*Familial Chylomicronemia Syndrome*" (FCS). PubMed & Google Scholar (GS) returned zero hits. The second query was "*Familial Chylomicronemia Syndrome mechanism of disease*", PubMed returned 1 hit and GS zero hits. For both queries, spot.my returned over 100 results each which identified atherosclerosis, multi-organ dysfunction and high co-incidence of Parkinson's and Alzheimer's.

2 Word frequency cloud suggests terms to improve the existing query

3 Disease factors implicated in a rare disease (*Acute Intermittent Porphyria)* were absent from Google and PubMed results

Deep semantic searching is superior to limited keyword searches, particularly for rare diseases or obscure news articles. Semantic results can provide clues to inform future areas of detailed exploration. Therefore, deep semantic search could be extraordinarily useful for intelligence-related search.

4 Keyword searching requires precise wordsmithing – semantic search is more forgiving and flexible

As the length of the query increased, PubMed ceased to work (n>50 characters) and Google Scholar became increasingly inaccurate (since it searched on so many keywords in parallel) until it ceased to work by n=256 characters. In contrast, spot.my offered search with greater context and the quality of search improved as more context was added to the search query.

Our Research Interests

- Simultaneous multilingual semantic extraction and mapping
- Deep personalization with individual indexes for each analyst
- Visualization and visual search of concept clusters
- Automated ontology generation for intelligence domains

Serdar Uckun, MD, PhD

Chief Operating Officer

RowAnalytics Inc.

serdar@rowanalytics.com

650-743-7192

www.rowanalytics.com

One Broadway, 14th Floor

Cambridge MA 02142