CAUSALITY LINK

Analytics Controlled Narratives with SRAG

October 25th, 2023

Olav Laudy, Eric Jensen, Pierre Haren

Contents

Summary

In today's information-saturated financial world, surfacing in real time the nuggets that make a difference in the performance of companies, industries and economies is essential for the asset management community. Large Language Models (LLM) summarization is both unable to support real-time information and prone to hallucinations, leading to the emergence of Retrieval Augmented Generation (RAG), which focuses the attention of the LLM on specific fragments of recent information.

Most RAG systems leverage statistical distance between concepts through embeddings. We present here a different approach: Structured RAG (SRAG), which uses a symbolic approach to convert financial articles into readable structured data that guides the selection of text fragments for the prompt. The availability of this structure enables the detection of insights leveraging topic frequency, sentiment trends and intricate causal networks, producing Analytics Controlled Narratives that are accurate, relevant, and rooted in analyzed data.

causalitylink.com

Introduction

In our digital age, the daily influx of information is overwhelming, leaving individuals and corporations alike grappling to discern the vital from the trivial. The modern world presents us with an irony: while data is abundant, the time and capability to extract meaningful insights from this deluge are scarce. Consider the finance sector, for instance. A myriad of articles, reports and analyses provide a daily pulse on companies, industries, and macroeconomic landscapes. But who has the time or capability to read and synthesize it all, especially when the stakes are high and decisions need to be prompt and informed?

The Causality Link solution emerged from this very conundrum. Our aim was straightforward, albeit challenging: to streamline the vast array of information into digestible, actionable insights. Harnessing the power of artificial intelligence, we've developed a research platform that does more than merely scan and summarize. It taps into the collective intelligence of tens of thousands of news articles each day, gauging crowd consensus. This enables a nuanced understanding of the perception shifts around companies, sectors, and macroeconomic indicators.

To ensure accuracy and contextual relevance, we developed an approach that merges the best of generative AI with symbolic AI, ensuring that the output is not just data-rich, but also contextually sound and analytically robust – an Analytics Controlled Narrative.

In the chapters that follow, we'll delve into this methodology: Structured Retrieval Augmented Generation, or SRAG. We'll outline how analytics can be constructed using SRAG and eventually lead you into the world of Analytics Controlled Narrative, where structured analytics drives the story, ensuring relevance, precision, and value.

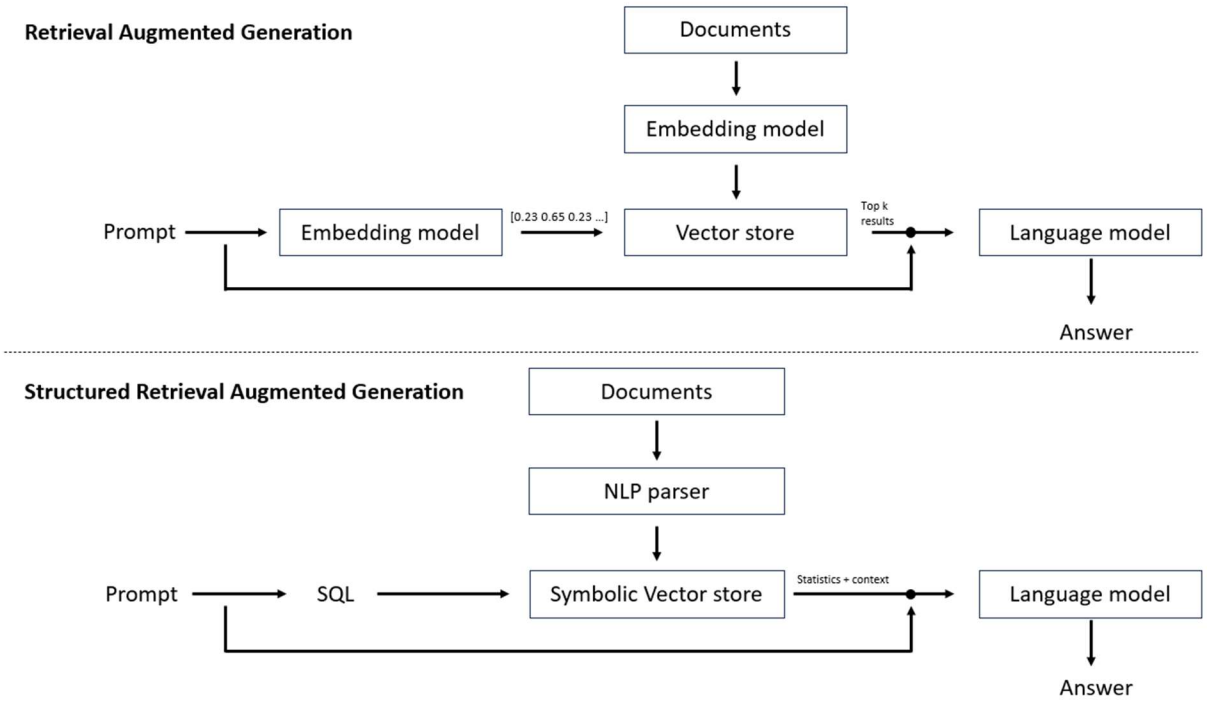
As we progress, we hope to illuminate the promise and potential of a world where AI doesn't just simplify our informational landscape, but also makes it more meaningful, guiding us toward informed decisions in an increasingly complex world.

Structured RAG (SRAG)

In the realm of artificial intelligence, a deep understanding of how data is retrieved, processed and presented is crucial. Retrieval-Augmented Generation [Lewis et. Al, 2020], or RAG, emerges as one such method that seamlessly bridges the gap between information retrieval and content generation.

Here is the definition of RAG according to Cohesity in 2023: "In a RAG-based AI system, a retrieval model is used to find relevant information from existing information sources while the generative model takes the retrieved information, synthesizes all the data and shapes it into a coherent and contextually appropriate response."

Figure 1: RAG and SRAG architecture, after Perez L. (2023)



Today, RAG implementations use a vector database as the retrieval component. Here is a breakdown of the components and the process:

Prompt: This is the initial input or question provided to the system. It serves as the starting point for the retrieval and generation process.

Embedding Model: The prompt is passed to an embedding model, which converts the text of the prompt into a high-dimensional vector. This vector is a numerical representation of the prompt that captures its semantic meaning.

Vector Store: This is a storage system containing vectors (high-dimensional representations) of all the documents in the system's knowledge base. When the prompt is embedded, its vector is compared to the other vectors in the vector store to determine which documents are most relevant.

Documents: These represent the vast amount of text data or knowledge the system has. Each document in this collection has its own vector representation stored in the vector store, generated using the same embedding model.

Top k Results: After comparing the prompt vector with the vectors in the vector store, the system retrieves the top-k most relevant documents. The value of "k" can vary, but it typically represents a small number of the most closely matched documents.

Language Model: The selected top-k documents, along with the original prompt, are then passed to a language model. The language model uses this information to generate a coherent and contextually appropriate answer. By using the retrieved documents as context, the language model can provide more accurate and information-rich responses.

In sum, the RAG process involves:

- Converting the prompt to a vector using an embedding model.
- Retrieving the most relevant documents based on the similarity of their vectors to the prompt vector.
- Using these documents as context to generate an answer with a language model.

The RAG model, while innovative, has inherent limitations. A primary issue is that vector databases, central to RAG, lack a built-in concept of time. As a result, users must manually add metadata to fine-tune results based on embedding similarities. This manual intervention can lead to inefficiencies and inaccuracies.

Also, current advancements in the field indicate that simple retrieval isn't always optimal. More complex retrieval methods, such as multi-step approaches, are being explored [Asai A., et al, (2023), Siriwardhana, S., et al. (2021)].

Another challenge for the RAG model is versioning. Data encoded with one embedding model version requires the same version for decoding. This means that a single prompt might need to be processed across multiple model versions, adding complexity to the retrieval process.

Given these challenges, it's clear that alternative methods to the usage of embeddings can and should be considered. At Causality Link, we use something different: a symbolic store.

Symbolic Database

Think of a symbolic database as an evolved tagging model. Instead of mere abstract representations of texts as in a vector database, a symbolic database provides references about detected patterns stored via columns that are both human-readable and interpretative. It's as if most of the fundamental ideas of all the documents were extracted in readable structured format and stored in a database, as opposed to

a vector database, where all the chunks are represented as large, incomprehensible vectors of real numbers.

Symbolic databases have many advantages:

Ontological Agreement: The database is not a mere collection of text references. It's structured around an ontology, ensuring common understanding and agreement on term definitions. This common ground ensures consistency and minimizes ambiguities across documents. The ontology represents static knowledge about the world that does not have to be “discovered” through a machine-learning process, thereby reducing considerably the costs of such a system.

Harnessing Traditional Technology: Despite our fascination with the potential of LLMs, we find it always beneficial to leverage established technologies. A symbolic database, by virtue of its structure, can easily integrate with traditional database technology.

Structured Data Output: Structured data offers granularity and dimensionality. Users can analyze specific topics in detail or aggregate data across dimensions like time, KPI, geography and industry. This creates comprehensive and targeted analytics, essential for informed decision-making.

Numerical Measures: In a symbolic database, while markers serve as descriptors, it's apt to term stored numbers as numerical "measures." These measures provide quantifiable insights, like the gap between article publication and forecast dates, enabling precise analytics.

The symbolic database enhances the RAG method into an SRAG method. Instead of broad similarity searches across text, this database allows for much more precise queries. This is especially useful when seeking specific evidence for financial inquiries. The result is targeted retrievals based on exacted citations from the database.

In the next chapter, we will examine how SRAG processes raw data into clear insights.

Analytics with Symbolic Database

The real potential of the symbolic database becomes apparent when paired with advanced analytics. Rather than just serving as a retrieval and generation system, it acts as a medium to transform dense financial articles into structured, actionable insights.

The NLP parser, the symbolic counterpart to the embedding model, is crucial in this transformation. It processes dense financial articles, converting them into understandable structured data. This data is optimized for statistical extraction and maintains connections to its original textual source. Ensuring that analytical data is traceable to its origin is vital for validation and credibility.

Unlocking a Spectrum of Analytics:

At the heart of analytics lies the power to aggregate data and explore it in diverse ways. With the symbolic database's tags and measures, users can seamlessly navigate a dashboard, zooming in on specific details or zooming out for a broader view. Filters can be applied to zoom in on particular topics, sectors or timeframes. Moreover, by leveraging the underlying ontology, the data can be rolled up to provide summarized insights or drilled down to reveal granular details. This dynamic exploration is a cornerstone of analytics, offering users the flexibility to shape their data narratives.

Ordering by Count/Frequency: Frequency evaluation is a foundational form of analysis. It helps identify dominant topics, frequently mentioned companies or sectors in the spotlight, painting a picture of current discussions.

Topic Acceleration: Beyond frequency, it's vital to understand momentum. Which topics are gaining traction over a short period? This can be a precursor to identifying emerging trends or issues before they become mainstream.

Sentiment Over Time: The pulse of (trend) sentiment, whether positive, negative or neutral, provides invaluable insight into market perception. By tracking sentiment chronologically, one can gauge how perceptions about a company or sector are evolving.

Sentiment by Past/Forecast: Distinguishing between retrospective and forward-looking sentiments can offer distinct perspectives. While the former provides a performance analysis, the latter gives a sense of market expectations and predictions.

Causal Networks: Delving deeper, Causality Link's technology automatically extracts causal explanations from text. This enables the mapping of intricate webs of cause-and-effect relationships, bringing forth a more profound understanding of interconnected market events.

Bayesian Networks: Continuing from causal relations, the data can be processed further to create valid probabilistic models [Laudy O. et al. (2022)]. These Bayesian networks, built on the foundation of the extracted causal networks, provide probabilistic insights into potential market outcomes.

Convergence of Structure and Insight:

The capabilities unlocked by SRAG go beyond individual insights – they form interconnected components of a comprehensive analytical framework. When these insights are pieced together, they present a unified, panoramic view of market dynamics, sentiments and their complex interplays. This isn't just about quantity; it's about the depth, breadth and interconnectivity of the information.

The power of this structured analytical approach lies in its embodiment of collective intelligence. By incorporating a diverse range of data sources and perspectives, it captures the multifaceted nature of market dynamics. This diversity ensures that the analysis is robust and comprehensive, minimizing blind spots and biases.

In the forthcoming chapter, we'll bridge the gap between these analytics and narrative construction. We'll explain how these structured analytics drive meaningful, precise and contextually rich narratives. The aim is clear: to move beyond mere data to deliver narratives that resonate, inform and guide.

Analytics Controlled Narratives Using SRAG

We now introduce the concept of Analytics Controlled Narratives: what is it and why is it revolutionary?

The Limitations of Large Language Models (LLMs)

We must first comprehend some of the constraints of typical LLMs.

While LLMs are marvels in data synthesis and generation, they're not infallible. In a concise format, LLMs may sometimes generate conclusions that veer off course, producing either generic or irrelevant content. They might grasp the "story" but can falter in comprehending the overarching "narrative." This is especially evident when the task demands tracking and aggregating collective intelligence from an ocean of sources.

Furthermore, LLMs are trained on a fixed dataset. Their knowledge base, especially concerning real-time financial events, becomes dated over time. This inherent limitation amplifies the importance of RAG, which allows for real-time data retrieval from current databases, ensuring relevant and up-to-date content generation [Zhang P., et. al. (2023)].

However, RAGs provide only statistically similar context to what the prompt asked for, and while they solve for the timeliness of the information provided to the LLM, they also suffer from their statistical approach in selecting relevant content.

Analytics controlled narratives emerge as the solution to this challenge. At its core, this method employs database statistics derived from SRAG as a guiding framework for the generative AI component. Instead of allowing unfettered freedom, the process offers structured guidelines, ensuring the narrative remains relevant, precise and reflective of the analyzed data.

For instance, if 80% of forecasts regarding an economic entity are positive, this statistic can be channeled to the LLM. The instruction then becomes clear: craft a narrative (built from the underlying quotes, while keeping references intact) that aligns with this overwhelmingly positive sentiment. A traditional RAG would arbitrarily find statements that are similar to the question instead of synthesizing all the evidence.

The SRAG prompt to the LLM, on the other hand, serves as a blueprint, ensuring the AI understands the context and adheres to rules derived from wisdom-of-crowd insights. It's not about merely generating content; it's about generating content that accurately reflects the analyzed data.

Inside the Analytics Controlled Narratives Prompt

Setting the Context: Every narrative starts with laying the groundwork. For example, "The subsequent analysis revolves around the fundamental characteristics of company XYZ."

Rules of the Statistics: Here, we define how the statistical data will be presented and interpreted, such as, "Below, each factor is presented with its overarching past/future sentiment interpretation, supported by quotes from relevant news articles."

Objective: Clear-cut guidelines ensure the narrative remains on track. An example could be, "Summarize the findings in a maximum of five paragraphs, focusing on specific evidence rather than generalities."

Data Organization: The narrative must present evidence in a structured manner, like, "Each piece of textual evidence starts with its date and the specific KPI mentioned, followed by the quote and its source."

Referencing: Credibility demands traceability. For example, “Each claim or piece of evidence must be anchored with references, ensuring transparency and trustworthiness.”

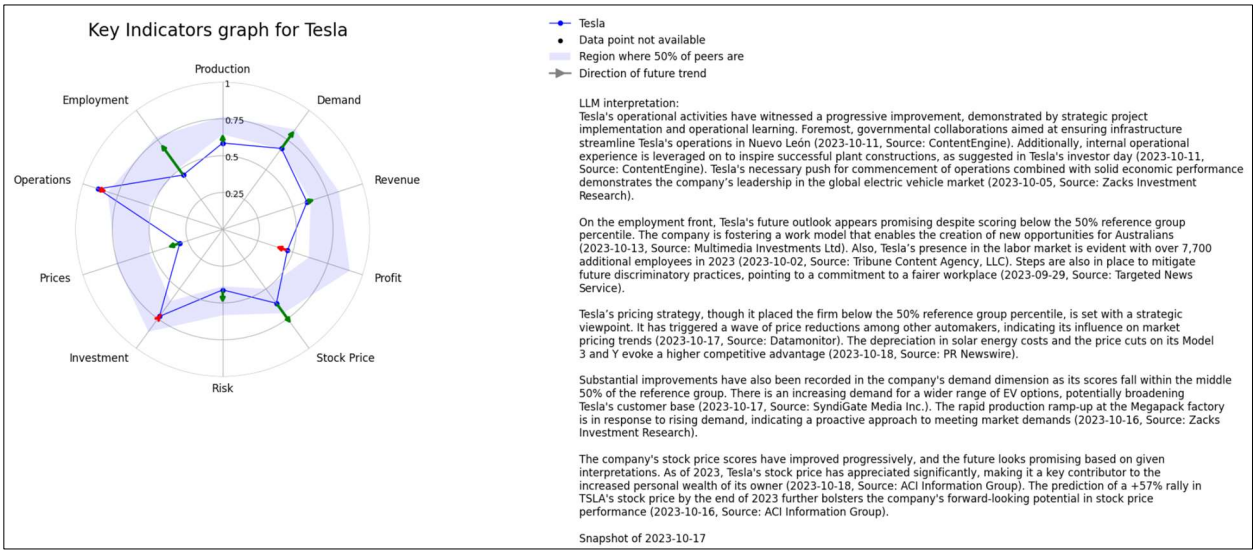
Context Delivery: An organized collection of quotes representing the underlying source of the statistics at hand.

Example 1: Tesla KPI Summaries

Practical examples can showcase the potential of Analytics Controlled Narratives and the fusion between structured analytics and narrative precision. One such example is presented here, focusing on electric vehicle giant Tesla.

Note that these examples use data available as of October 17, 2023, and that Tesla subsequently announced its results. During the presentation, CEO Elon Musk made surprising negative statements on the impact of high interest rates on future demand, as well as more evasive statements about the creation of the Mexican Gigafactory. This negatively impacted Tesla’s stock price in the short term, contrary to the end-of-year expectations prior to the earnings call.

Figure 2: Example 1: Summarize



This Key Indicators graph for Tesla offers a visual representation of various metrics pertinent to the company as of October 17, 2023. These metrics range from Production to Stock Price, encompassing essential facets like Revenue, Operations and Risk. Each metric is evaluated against a normalized scale, providing comparative insight into Tesla's performance relative to its peers in the industry. The scale ranges from 0-1, with 0 being full consensus of the KPI declining, 0.5 being split consensus and 1 being full consensus of the KPI increasing.

As illustrated in the graph, Tesla demonstrates a noteworthy trajectory, particularly in the areas of Production, Operations and Employment. These high scores are emblematic of the company's ongoing progress and strategic advancements in the market.

The LLM interpretation alongside the graph uses quotes from the majority sentiment to dive deeper into these achievements. For instance, Tesla's notable progress in its operations is highlighted, marked by milestones such as its burgeoning operations in Nuevo León, Mexico. These references can then be links to the source articles.

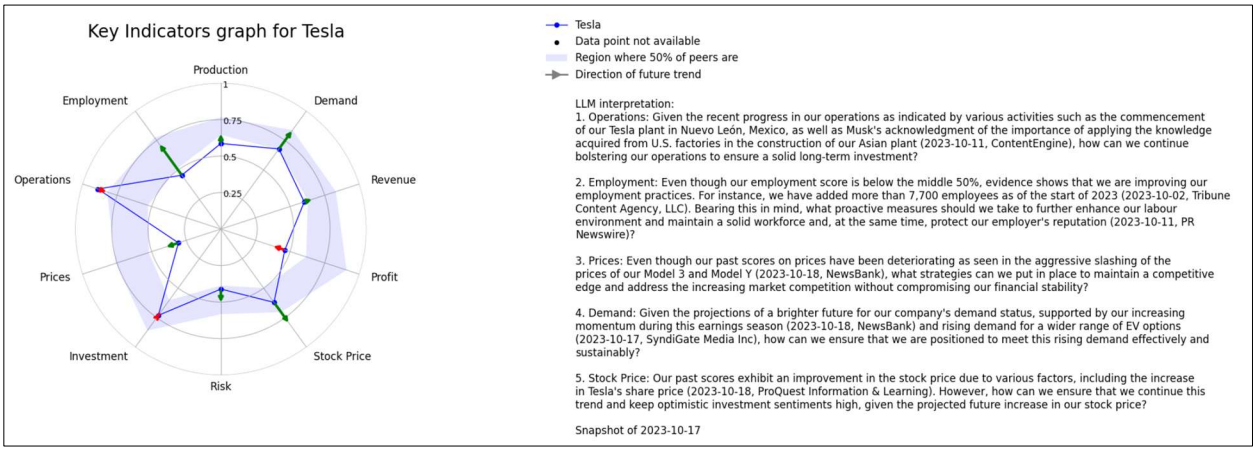
The summaries are presented in descending order of deviance (the absolute distance from 0.5 plus the size of the future-past arrow). This aims to mimic how a dashboard user would want to see the topics ranked by importance.

In summary, this example demonstrates how an Analytics Controlled Narrative, bolstered by structured data, can automatically provide a holistic overview of a company's performance at any given date. Such narratives, grounded in fact and contextual understanding, are invaluable for stakeholders aiming for an informed perspective.

Example 2: Tesla Independent Board Member Questions

The second example shifts our attention to possible questions from an inquisitive independent board member concerning Tesla's trajectory and decision-making strategies. While the provided data is the same, the instructions now ask to create no more than five thoughtful and critical questions regarding the presented information. Independent board members do not always have the time to stay abreast of all the current news about a company, and this cheat sheet can prove invaluable to performing their supervisory duties.

Figure 3: Example 2: Tesla Independent Board Members' questions



Conclusion

Analytics Controlled Narratives, driven by SRAG analytics, represent a paradigm shift: a future where AI-generated narratives are not only rich in data but also in context and relevance, where every story is not just told, but told meaningfully, rooted in a foundation of structured analytics.

As we progress into this new era, the potential for more informed decision-making, clearer insights and more profound understanding awaits.

References

Asai A., et. al. (2023) Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. https://arxiv.org/abs/2310.11511

Cohesity (2023) https://www.cohesity.com/glossary/retrieval-augmented-generation-rag/

Laudy, O., Denev, A., and Ginsberg, A. (2022) Building Probabilistic Causal Models Using Collective Intelligence. Journal of Financial Data Science. DOI 10.3905/jfds.2022.1.091

Lewis, P., et. al. (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. https://arxiv.org/abs/2005.11401

Perez, Luis L. (2023) Taming the Parrot: A tutorial on Retrieval Augmented Generation, slide 3. https://www.linkedin.com/posts/abu-dhabi-machine-learning_luis-talk-on-rag-taming-the-parrot-activity-7116491348836311040-mZx1

Siriwaardhana, S., et. al. (2021) Fine-tune the Entire RAG Architecture (including DPR retriever) for Question-Answering. https://arxiv.org/abs/2106.11517

Zhang P., et. al. (2023) Retrieve Anything To Augment Large Language Models. https://arxiv.org/abs/2310.07554