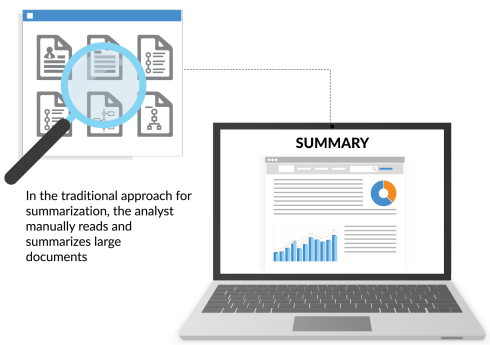
Summarizing Large Corpus of Documents using AI and Knowledge Graphs for Better Context

—Anshul Ujlayan, Ph.D. | Mastech InfoTrellis

In today's world of information overload, information processing technology is unarguably thriving. Retrieving and filtering documents from a vast volume of text documents is not too difficult for most organizations. However, reading a large number of documents takes a lot of time and effort, and it's a problem that calls for a systematic approach to solve. Organizations need an intelligent system that can correctly summarize the documents' context. A Knowledge Graph generated based on inputs from a system-generated summary of a large set of documents is an apt solution. It provides a quick view of the summary for business use with a detailed visualization.

The traditional approaches to Summarization

The information database of IT companies does not necessarily follow a uniform format and, therefore, must be pre-processed for business use. The pre-processing of semi-structured and unstructured data in organizations is gaining popularity as a practice with many approaches to capture the context. As it's still challenging for enterprises to manage and store knowledge for real-time business use, summarizing such information from a considerable amount of text documents is essential to speed up the process. The industry's traditional approach to managing a lot of information is to summarize content using manual effort from subject matter experts. The experts read and understand the complete information and summarize that in a short form without losing the context.



Contents

Moreover, there are many questions about the traditional approaches of the summarization and visualization of Knowledge Graph.

1. How accurate is the generated summary?
2. Does the summary capture the full context of the large document?
3. How can we optimize the time to summarize the large document?
4. How can we prepare the interactive visualization for the Knowledge Graph?

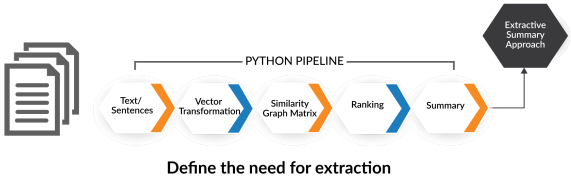
Organizations spend a lot of time and effort and use advanced tools and technologies to answer the above questions. When given a large amount of text data, a user can create a summary and visualize it in the form of a Knowledge Graph that best captures the large text documents' facts while adhering to a set of constraints. This automatically sheds light on a lot of unutilized information in the organization that can now be used for business purposes. The Knowledge Graph visualization can provide deeper insights into the connection of different documents in the form of connected nodes with a relationship as edges.

The Summarization Approach with Machine Learning and Knowledge Graphs

Summarization is the process of extracting valuable knowledge in less time in the form of a synopsis from a large text content without losing the context. Document summaries, for example, provide a brief overview of the material until a user digs deeper. Since entity descriptions, including web pages or text, are extensive, they must be presented logically. Organizations around the globe spend a lot of money, in addition to resources, to prepare a good summary from a large text in a short time. Our proposed Text Summarization Solution can help them achieve their goal with greater accuracy. There are two main approaches to summarization:

Extractive Summarization:

Extractive text summarization extracts phrases and sentences from the source text to generate a summary. This approach uses the technique of ranking the importance of phrases and select only those most important to the source's context.

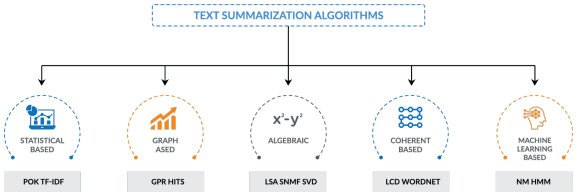


Define the need for extraction

Abstractive Summarization:

Abstractive text summarization is about creating completely new phrases and sentences to capture the context of the source document. Though this is a much more complex strategy, it is one that humans can use. The content of the source text is selected and compressed using traditional methods in this approach.

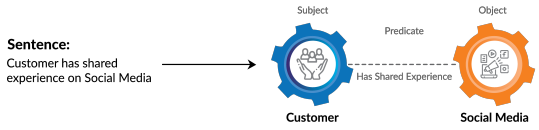
There are various algorithms available for extractive and abstractive summarization, which are shown below:



The documents are represented as a linked graph by methods that are inspired by the PageRank algorithm. The vertices of the graph are sentences, and the edges between them show how similar the two sentences are. A common method for connecting two vertices is to compare the similarity of two sentences and link them if the similarity is greater than a certain threshold. Summarization is modeled as a classification problem in machine learning approaches.

Knowledge Graph

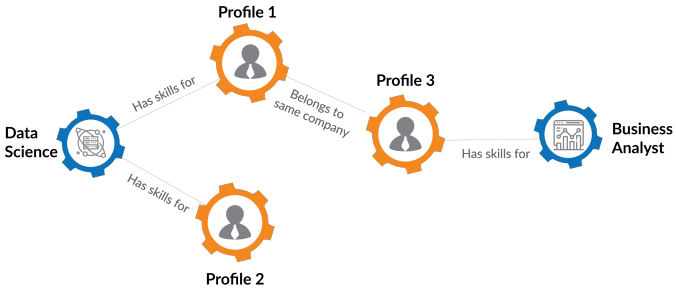
A Knowledge Graph is made up of a collection of individuals, relationships, and triples. A triple is a set of three entities linked by a relationship - subject, predicate, and object. Facts in a Knowledge Graph are known as triples in natural language. According to graph theory, triples are edges that connect pairs of entity nodes. Knowledge Graphs are created after summarization by capturing domain-specific knowledge as a data layer and adding rich and explicit semantics to infer additional knowledge. We can define the Knowledge Graph as a set of connected nodes or entities with a defined relationship as edges. There can be multiple relations existing between various nodes in the case of a large text corpus. For example:



The data in the above example can be stored in a triplet format i.e.

subject, predicate, object = Customer, HasSharedExperience, SocialMedia

There can be multiple relations between various nodes in the case of a large text corpus. Here is an example of the recruitment industry.



In the figure above, several candidates' profiles may have the same skill sets and experience, and some may also have worked in the same organization in the past. Knowledge Graphs can bring out such relationships in a visual manner with ease.

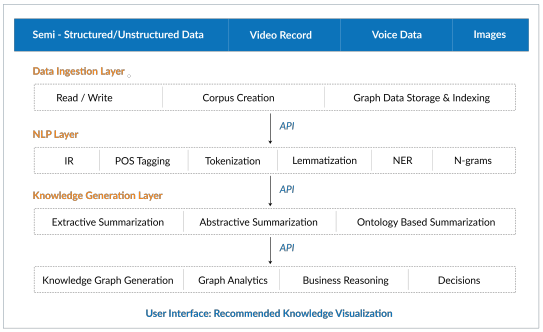
Large information graphs are generally used in the scientific community, such as encyclopedic datasets like DBpedia, Freebase, FIBO, and Yago. Typically, they are made up of millions of entities and billions of facts that describe them. Artificial intelligence and machine learning approaches leverage advanced technologies like linked open data commercial information graphs to deliver good results that help a business.

Business Use of Summarization and Knowledge Graph Generation

Summarizing a large amount of information helps save the time spent reading and understanding the subject matter. In large organizations, summarization reduces the time spent searching historical documents like legal documents, web contents, business email and documents, news, and scientific articles. Intelligent summarization and Knowledge Graph generation directly address the issues arising from the time-consuming process by summarizing, managing, and storing many documents in an organization. Automated Knowledge Graph generation and visualization give businesses real-time benefits to use historical documents to plan and strategize efficiently.

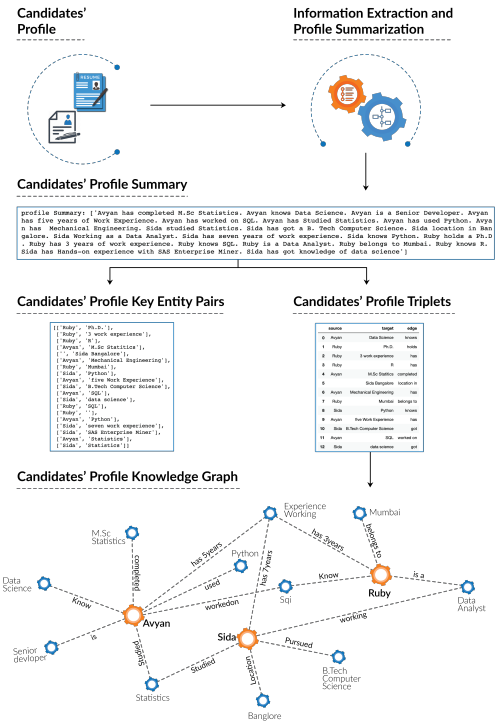
Mastech InfoTrellis Cutting Edge Solution

The machine learning and ontology-based solution for summarization and Knowledge Graph generation from Mastech InfoTrellis can bring about a new organizational experience. The solution, which works efficiently on large text documents, is built by customizing several machine learning algorithms.



Summarization in Recruitment to Reduce Candidate Screening Time

In the recruitment industry, recruiters are often forced to spend a lot of time scrutinizing candidate profiles and understanding their profile attributes. The Text Summarization Solution can generate a quick summary and visualize the Knowledge Graph of all the candidates instantly by finding connections between the candidates' profiles, significantly reducing the time spent screening the profiles. The system ingests all the candidates' profiles and performs a python-based data pre-processing using natural language processing. Post data normalization is then processed for profile summarization, and the output of profile summarization is used to generate a Knowledge Graph of the candidates' profile as shown below:



In the illustrative example above, the solution uses natural language processing to extract all the relevant information from the candidates' profiles based on human resource ontology. After the information extraction, the entire raw text corpus is transformed into a normalized form for the implementation of customized machine learning algorithms for summarization. The system-generated profile summary is then used to identify the triplets (i.e., subject, predicate, object) for individual sentences in the profile. These triplets are further used to generate the knowledge graph of the candidates' profiles. The knowledge graph shown in the illustration provides a clear view of the candidate's profile attributes as well as the similarity in the skillsets and the candidates' profile linkage.

Conclusion

In contrast to "traditional" approaches, our solution leverages a new approach to extract, summarize, and visualize the underlying information. We leverage AI/ML to automate the process of information summarization and Knowledge Graph visualization for a large set of documents. This modern approach helps enterprises process a large number of business documents and get a quick summary and knowledge view without having to screen long historical text documents manually. The Knowledge Graph extraction and visualization provide insights to review the individual or multiple documents and identify how they are connected. As technologies evolve, enterprises are compelled to explore advanced machine learning and deep learning algorithms such as natural language generation and abstractive summarization for a robust solution. According to related research literature, multiple machine learning techniques used together can produce a better hybrid approach for Knowledge Graph generation.

Author

Anshul brings 14+ years of experience in leveraging ML/AI to solve business problems. He holds a BS and MS in Applied Statistics, and a PhD in Management.

About

Mastech InfoTrellis partners with enterprises to help them achieve their business objectives by leveraging the power of data to derive deep, analytical insights about their business and its operations. We accelerate business velocity, minimize costs, and drastically improve corporate resiliency through personalized, process-oriented programs, consisting of strategy, data management (including master data management), business intelligence and reporting, data engineering, predictive analytics, and advanced analytics. Part of the NYSE-listed, $193.6M, digital transformation IT services company, Mastech Digital; we drive businesses forward around the world, with offices spread across the US, Canada, India, Singapore, UK, and Ireland.