GA-Tag: Data Enrichment with an Automatic Tagging System Utilizing Large Language Models

Genki Kusano
NEC Corporation
Tokyo, Japan
g-kusano@nec.com

2024 IEEE 40th International Conference on Data Engineering (ICDE) | 979-8-3503-1715-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDE60146.2024.00412

***Abstract*—Data quality is widely recognized as being directly linked to the quality of analysis results. In this study, we introduce a tagging method that simplifies the handling of extensive data and facilitates the rapid search and extraction of relevant information. Traditional methods that search for and integrate related data from external sources to enrich input data often fail to guarantee the acquisition of desirable information for all data sets. However, the recent advancement of Large Language Models (LLMs) enables the prediction of characteristics of input data, even in the absence of relevant data. In this paper, we present the *Generated and Aggregated Tag* (GA-Tag), a system that employs LLMs to automatically assign appropriate tags to data and is equipped with an aggregation mechanism to manage tag diversity effectively. The adoption of GA-Tag is anticipated to enhance data analysis and management quality and efficiency, optimize monetary and time costs, and potentially bolster business intelligence and decision-making processes.**

***Index Terms*—Tagging System, Large Language Models, Data Enrichment, Clustering**

I. Introduction

As the concept "Garbage in, garbage out" indicates, if the input data lacks quality, the resulting outputs will also be insufficient, no matter how advanced the data analysis methods are. This phrase emphasizes the pivotal role of data quality in effectively utilizing database management systems. To avoid this phenomena, a data enrichment method [1] has been proposed that involve searching for related data from external data sources and integrating it. However, there is no guarantee of always obtaining the desired information. For example, in the context of company data, when attempting to enrich the database with user characteristics, external sources do not always contain sufficient information for all companies.

Even if no data source contains answer information, recent advancements in Large Language Models (LLMs) [2]–[4] have made it possible to predict such characteristics, which is achieved by utilizing the knowledge embedded within the LLM and inferring various relationships.

In this research, we explore *tagging* as a method for data enrichment. Tagging simplifies handling extensive data, allowing for the rapid search and extraction of relevant information, as commonly used on social media and news websites. Figure 1 depicts a scenario where tags are assigned to each company, and a column of tags is newly added to the company table. With the enriched table with tags, it becomes possible to obtain a variety of insightful analyses that could not be achieved with untagged input data.

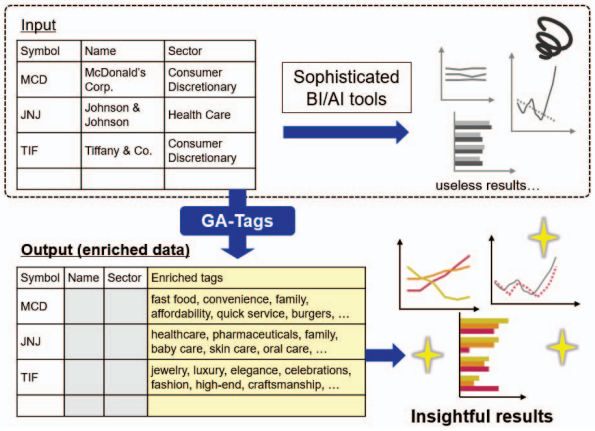

Fig. 1. Even when it is difficult to obtain valuable results with the input data alone, enriching the data with GA-Tag system enables the extraction of insightful insights.

In this demonstration, we present *Generated and Aggregated Tag* (GA-Tag), a system for automatic tagging utilizing LLMs. GA-Tag first generates tags for each data entry with the assistance of an LLM. It takes advantage of the model's ability to simplify tagging, removing the need for additional data collection or training models. Although employing LLMs facilitates tagging, the diversity in tag expressions can complicate subsequent data analysis tasks, such as grouping data with identical tags or creating tag histograms. This problem occurs because tags with similar meanings can be expressed with different words. GA-Tag resolves this issue of excessive tag variety by incorporating an aggregation mechanism to group similar tags together. It systematically reduces the total count of tags, ensuring that data analysis remains efficient and insightful.

In recent developments involving LLMs, users increasingly pay attention to the costs associated with processing through these models. Since GA-Tag produces compact tag expressions that require fewer tokens than lengthy texts and incorporates a batch processing mechanism, it reduces monetary and time costs, ensuring financial and operational efficiency.

II. System Overview

In this section, we will explain GA-Tag system, where the back end comprises two processes: generating tags for each

2375-026X/24/$31.00 ©2024 IEEE
DOI 10.1109/ICDE60146.2024.00412

Authorized licensed use limited to: East China Normal University. Downloaded on September 21,2024 at 06:53:58 UTC from IEEE Xplore. Restrictions apply.

data entry and aggregating similar tags. In the subsection on the front end, we will discuss how users can effectively operate GA-Tag system and what types of customization are possible.

A. Back end

1) Tag Generation Process: Let $R = \{r_i\}_{i=1}^{n}$ be a table with n rows and m columns. For each row $r_i = \{(a_j, v_{ij})\}_{j=1}^{m}$, a_j represents the j-th column name in R and v_{ij} denotes the element at the i-th row and j-th column. In the context of this research, each row of the table is treated as data for enrichment. In the case of the example in Figure 1, the row $r = \{(\text{Name}, \text{Nestlé.}), (\text{Sector}, \text{Consumer Staples})\}$ represents the input data for GA-Tag.

To perform tagging on the target data r, we utilize LLM[1]. Since r represents a row of a table, we transform it into text using the designated prompt as shown in Figure 2.

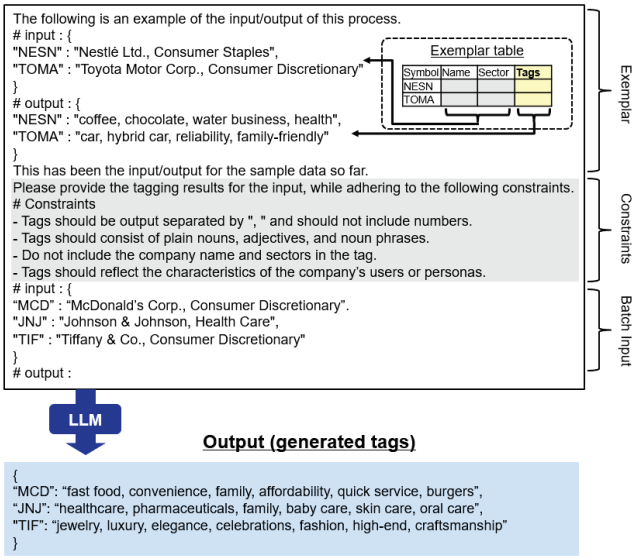

Fig. 2. Prompt for textualized input data.

The prompt for tag generation incorporates three key practices as listed below: [Exemplar] It guides the LLM in generating tags in the intended direction with few-shot samples. In Figure 2, an exemplar table with the desired tags added to the "Tags" column is prepared and loaded to the prompt. [Constraints] These refer to specific conditions imposed on LLMs, including the output format and the characteristics of tags. In the example of Figure 2, the top two are generic, while the bottom two are domain-specific to the data. [Batch Input] It feeds the input data to the system, which supports not just individual rows but also batch processing capable of managing multiple rows simultaneously.

As seen in the output of Figure 2, the generated tags are represented by concise words, enabling users to quickly understand the input data owing to the practical constraints added in the prompt. However, tags generated by LLMs are highly diverse, resulting in different data rarely sharing common tags, which makes data analysis difficult. This issue will be addressed in the subsequent section.

2) Tag Aggregation Process: We introduce a tag aggregation process that replaces similar tags with a representative term. For example, tags "apple", "orange", and "grape" can appropriately be categorized under a common label, such as "fruit". In this context, the tags produced during the tag generation process are referred to as "generated tags"; those replaced with representative terms are referred to as "aggregated tags". To create aggregated tags from given generated tags, we implement a two-step process:

1) Cluster tags to group semantically similar ones together.
2) Assign a representative label to each cluster.

Clustering: One standard method for grouping semantically similar words is to generate embeddings for them and then apply clustering algorithms based on these embedding vectors. We here adopt the approach introduced in [5], where tags are transformed into embedding vectors using RoBERTa [6][2] and grouped by using Agglomerative Hierarchical Clustering[3] on these vectors. In practice, embeddings like BERT are high-dimensional ($d = 100 \sim 10000$), and the hubness phenomenon may influence such high-dimensional data [7], which is also known as "curse of dimensionality". To mitigate this hubness phenomenon, we employ a method proposed by [8], subtracting the mean vector from the embedding vectors.

With regards to the embedding of tags, we designate a genre g of the tag and convert a tag t to a text "This $\{g\}$ has a tag of $\{t\}$" to be embedded. For instance, when t ="bank", the word itself could have various meanings like "riverside" or "financial institution". By setting as g ="company", the contextualized embedding will prioritize the interpretation of "bank" as a "financial institution".

Assigning representative labels: To represent each cluster with a short word, we input the prompt shown in Figure 3 into the LLM. Similar to the tag generation, we provide exemplars, impose constraints, and process the clusters in batches.

3) Output forms of GA-Tag: Once the tag aggregation process is completed, an aggregated tag a is assigned to each corresponding generated tag t. Let $C = \{(t, a) \mid t \in T, a \text{ is an aggregated tag of } t\}$ be a corresponding table between generated and aggregated tags, as shown in Figure 4. Using this correspondence C, we replace all generated tags with the corresponding aggregated tags. Finally, GA-Tag outputs an enriched table with a new column comprising generated and aggregated tags.

The enriched table can be displayed as a pivot table with rows of data entries and columns of tags or as an unpivot table. The pivot format suits function in BI tools or AI technologies, while the unpivot format benefits memory conservation. The choice between formats depends on the application.

[1]In this research, we used `gpt-3.5-turbo-0613` model as the LLM.

[2]Unlike the original paper [5] that used BERT for embedding, we chose to use RoBERTa, an improved version of BERT. The RoBERTa model was obtained from https://huggingface.co/roberta-large.

[3]https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

Authorized licensed use limited to: East China Normal University. Downloaded on September 21,2024 at 06:53:58 UTC from IEEE Xplore. Restrictions apply.

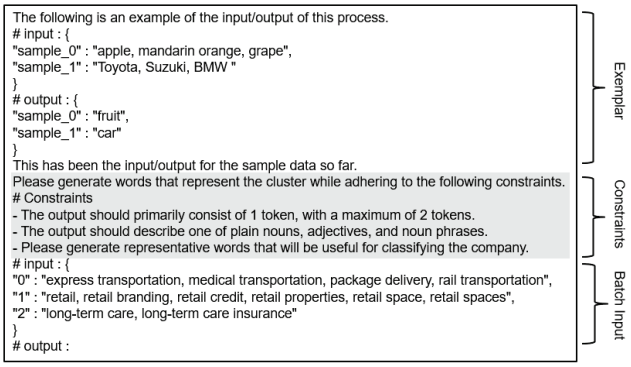

Fig. 3. Prompt for assigning representative labels to clusters.

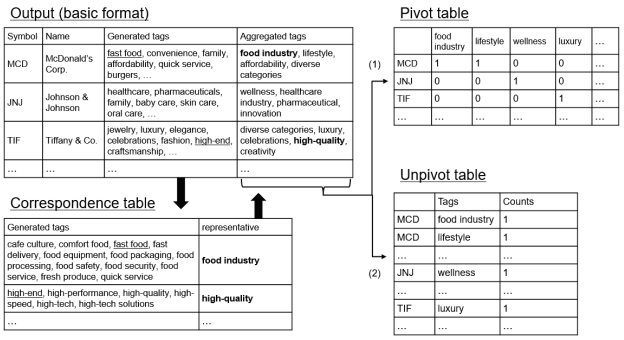

Fig. 4. The basic format of GA-Tag output (top left) and the pivot and unpivot table representations (right). Generated tags are converted into aggregated tags via the correspondence table (bottom left).

B. Front end

Figure 5 displays the settings screen[4] used for adjusting the back end components, with each configuration option briefly explained below: [I] Users input their data in a CSV table format. [S1] Users can upload an exemplar table. This table must have the same columns as the input data and include an additional column with examples of generated tags. If no exemplar table is uploaded, the sentences in the "Exemplar" part of Figure 2 will be omitted. [S2] Users can specify the genre of tags used when converting tags to text with RoBERTa. The default genre is "data", meaning a tag t will be embedded as the text "This data has a tag of $\{t\}$". [S3] Users can incorporate constraints that align with the sentences in the "Constraints" part in Figure 2. Clicking the "Add constraints" button can easily add additional constraints.

III. Demonstration

In this demonstration, we utilized a stock dataset comprising 500 companies from the S&P 500 [5] as an input for GA-Tag system. This dataset consists of a table with columns for

[4]This screen was made by using Streamlit https://streamlit.io/.

[5]https://github.com/datasets/s-and-p-500-companies/blob/main/data/constituents.csv

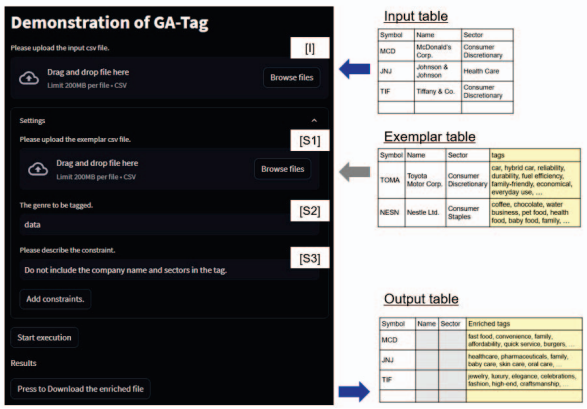

Fig. 5. Parameter Settings Screen

"Name" and "Sector", as depicted in Figure 1. To obtain the enriched S&P 500 data using GA-Tag, we follow these steps:

1) Upload the input table data.
2) Adjust several settings, as illustrated in Section II-B.
3) Execute GA-Tag system (Press the "Start execution" button in Figure 5)
4) Download the enriched file.

A. Scenario

In this section, we introduce a use case involving the application of enriched data. We follow a scenario where a data scientist examines a specific company, Marriott.

The data scientist decided to use the S&P 500 dataset to analyze the relative positioning of Marriott by comparing it with other companies. Initially, the original input data comprised only company names and sector columns. With over 80 companies in the same "Consumer Discretionary" sector as Marriott, it was challenging to identify all as direct competitors. To address this, the data scientist utilized GA-Tag system to assign tags to each company, enriching the dataset. A search for companies with "hospitality" and "luxury" tags narrowed the list to four, as shown in Figure 6. Among them, three operate in the luxury hotel sector and were recognized as Marriott's direct competitors.

Symbol	Name	Generated tags	Aggregated tags
MAR	Marriott Int'l.	hotels, travel, hospitality, luxury, business travel, family vacation	tourism industry, service industries, **hospitality**, **luxury**, business and travel, family activities
RCL	Royal Caribbean Cruises Ltd	cruise, travel, luxury, entertainment, vacation, relaxation, adventure, hospitality	tourism industry, service industries, **luxury**, vacation, lifestyle, **hospitality**
HLT	Hilton Worldwide Holdings Inc	hospitality, travel, luxury, comfort, service, vacation, business travel	service industries, **hospitality**, **luxury**, diverse industries, vacation, business and travel
MGM	MGM Resorts International	casino, resorts, entertainment, luxury, travel, hospitality	kitchen items, tourism industry, **luxury**, service industries, **hospitality**
HST	Host Hotels & Resorts	hospitality, hotels, travel, leisure, customer service, comfort, luxury, business travel, vacation, hospitality industry	**hospitality**, tourism industry, service industries, leisure activities, customer relations, diverse industries, **luxury**, business and travel, vacation

Fig. 6. Extracted data that contains "hospitality" and "luxury" tags.

Furthermore, these tags help conduct a comparative analysis between Marriott and its competitors. Using a pivot table with company names listed in rows and aggregated tags in columns,

Authorized licensed use limited to: East China Normal University. Downloaded on September 21,2024 at 06:53:58 UTC from IEEE Xplore. Restrictions apply.

the data scientist can easily understand the unique characteristics of each company at a glance. As shown in Figure 7, all competitors were assigned the "luxury", "hospitality", and "service industries" tags. On the other hand, only Marriott has the "family activities" tag, suggesting it could be a potential competitive advantage for the company. However, the absence of Marriott's "vacation" tag indicates a challenge in capturing the vacation-targeted customer segment.

	Name	hospitality	luxury	service industries	tourism industry	business and travel	vacation	diverse industries	customer relations	family activities	kitchen items	leisure activities	lifestyle
MAR	Marriott Int'l.	1	1	1	1	1	0	0	0	1	0	0	0
RCL	Royal Caribbean Cruises Ltd	1	1	1	1	0	1	0	0	0	0	0	1
HLT	Hilton Worldwide Holdings Inc	1	1	1	0	1	1	1	0	0	0	0	0
MGM	MGM Resorts International	1	1	1	1	0	0	0	0	0	1	0	0
HST	Host Hotels & Resorts	1	1	1	1	1	1	1	1	0	0	1	0

Fig. 7. Pivot table of Figure 6.

The applicability of GA-Tag is not restricted to company data. It can extend to various categories, including products (like food and daily necessities) and services (such as SaaS applications and insurance products). For specific individuals, tags about behavioral tendencies and preferences can be assigned based on their demographic attributes (age, gender, occupation, etc.), facilitating comprehensive customer analysis and targeted marketing strategies.

B. Performance

Here, we presented several statistics related to GA-Tag. The numbers in each table below show the average and standard deviation from three independent executions of GA-Tag.

C. Performance

Here, we presented several statistics related to GA-Tag. The numbers in each table below show the average and standard deviation from three independent executions of GA-Tag.

1) Tag: Table I shows statistics on the generated and aggregated tags, where the symbols $\sharp\text{tags}, \sharp\text{tags} \geq 5, P(\sharp\text{tags} \geq 5)$ represent the total number of tags, the count of tags appearing more than five times, and the proportion of such frequent tags, respectively.

	$\sharp$tags	$\sharp\text{tags} > 5$	$P(\sharp\text{tags} > 5)$
generated tags	1564.3 ± 8.34	156.3 ± 3.30	10.0 ± 0.24 (%)
aggregated tags	208.0 ± 8.04	125.7 ± 6.13	60.4 ± 2.23 (%)

TABLE I
SUMMARY OF THE NUMBER OF TAGS

Based on Table I, we observed that the unique number of aggregated tags is approximately seven times less than that of generated tags. The number of aggregated tags associated with more than five companies is nearly identical to that of generated tags. Consequently, the percentage of tags related to five or more companies has significantly increased from about 10% for generated tags to approximately 60% for aggregated tags, which indicates that aggregated tags are more effective for analyzing clusters of companies that share similar tags.

2) Costs: The API cost is a significant factor when developing services employing LLM. The API usage fee is mainly determined by the token length of both input and output text. Table II shows the token length statistics for prompts when the input file batch size is 20, as shown in Figure 2. As of the submission date of this paper[6], the total cost to use OpenAI API incurred for the tag generation process for 500 companies was 0.054 USD, while the tag aggregation process cost was 0.016 USD, making it relatively affordable.

		generation	aggregation
token	input	14995	9832.0 ± 178.3
	output	16198.0 ± 356.1	577.3 ± 23.2
cost (USD)	input	0.022 ± 0.0000	0.015 ± 0.0003
	output	0.032 ± 0.0007	0.001 ± 0.0000
time (second)		275.5 ± 11.3	69.0 ± 1.5

TABLE II
SUMMARY OF THE MONETARY AND TIME COST.

We also evaluated the time needed for processing via the OpenAI API. Table II shows that the tag generation and aggregation processes took 275 and 69 seconds, respectively. Since both tag generation and aggregation processes can parallel processing, executing the API in parallel can reduce the processing time proportionally to the number of parallels.

IV. Conclusion

In this paper, we introduced *GA-Tag*, a novel data enrichment method developed utilizing LLMs. GA-Tag possesses two core functionalities: automatic tag generation and tag aggregation. Additionally, it enables a reduction in both monetary and time costs, ensuring that data analysis remains efficient and insightful.

References

[1] Y. Dong and M. Oyamada, "Table enrichment system for machine learning," in *SIGIR*. ACM, 2022, pp. 3267–3271.

[2] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," in *USENIX Security Symposium*. USENIX Association, 2021, pp. 2633–2650.

[3] S. Ishihara, "Training data extraction from pre-trained language models: A survey," in *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*. ACL, 2023, pp. 260–275.

[4] V. Borisov, K. Seßler, T. Leemann, M. Pawelczyk, and G. Kasneci, "Language models are realistic tabular data generators," in *ICLR*. OpenReview.net, 2023.

[5] N. Reimers, B. Schiller, T. Beck, J. Daxenberger, C. Stab, and I. Gurevych, "Classification and clustering of arguments with contextualized word embeddings," in *ACL (1)*. Association for Computational Linguistics, 2019, pp. 567–578.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[7] M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Hubs in space: Popular nearest neighbors in high-dimensional data," *J. Mach. Learn. Res.*, vol. 11, pp. 2487–2531, 2010.

[8] I. Suzuki, K. Hara, M. Shimbo, M. Saerens, and K. Fukumizu, "Centering similarity measures to reduce hubs," in *EMNLP*. ACL, 2013, pp. 613–623.

[6]https://openai.com/pricing on October 23, 2023, OpenAI's `gpt-3.5-turbo-0613` charges 0.0015 United States dollars (USD) for 1000 input tokens and 0.002 USD for 1000 output tokens, respectively.

Authorized licensed use limited to: East China Normal University. Downloaded on September 21,2024 at 06:53:58 UTC from IEEE Xplore. Restrictions apply.