Proceedings of the International Conference on Industrial Engineering and Operations Management
Bangkok, Thailand, March 5-7, 2019

TAKA News Search Engine: A Proposed DEMO-based Performance Evaluation System

Tarek Fatyani
R&D Department, SMAGROUP LLC, Chiba, Japan
tarek@smagroup.co.jp

Zakaria Yahia
Department of Quality and Operations Management, University of Johannesburg, Johannesburg, South Africa
(on leave from Department of Mechanical Engineering, Fayoum University, Fayoum, Egypt)
zakariay@uj.ac.za, zakaria.yahia@fayoum.edu.eg

Abstract

In this paper, we develop an ontology-based performance evaluation system for a news search engine (called TAKA) using the enterprise engineering approach Design and Engineering Methodology for Organizations (DEMO). This paper has threefold objectives. First, to describe, understand and analyze the main processes in TAKA news search engine. For this objective, a conceptual DEMO model is developed for TAKA system. Second, to propose a performance evaluation system for TAKA system. For this objective, a set of measures is developed for each process. Third, to analyze the performance and to draw potential improvements for the TAKA news search engine. For this objective, each process is analyzed, and corrective decisions are to be considered. This paper contributes to the literature by presenting a performance evaluation system for the entire news search engine from user, enterprise and systems engineering point of views. On contrast, studies in the literature consider only the users' perspectives.

Keywords

News search system, news monitoring system, news search engine, design and engineering methodology for organizations, DEMO, performance evaluation system, key performance indicators, KPIs

1. Introduction

1.1. Background and Motivation

The number of online news sources has augmented vividly in the last decade. Today, an increasing number of people are reading news online which is mostly free and easily accessible. Furthermore, many kinds of information are sensitive to time, especially the case with news. In fact, the value of a news article depends much on its time of publication. A more important reason for online news popularity is that people can access breaking news articles as soon as they are posted. Furthermore, old news that may not be readily available from newspapers can be obtained.

However, the availability of large amount of online news overwhelms news-consumers. Moreover, because the enormous number of news sources, specialized news sites, and with the huge amount of online news; a news article of interest to a person may be posted in a newspaper web site unknown to him/her.

News search engines are introduced as a solution to those issues. They aim to deliver news in an effective (according to consumer needs) and efficient (effortless) manner. They equipped with a wide range of tools which facilitate browsing news. With search tool, they can return fresh and relevant news in response to a news query. Moreover, their filtering capabilities deliver latest news according to consumers' interest. Such systems enable news-consumers creating their own aggregated online newspapers to receive news from multiple sources. They allow users to search news articles from several news sources around the world from a single search system. Their inclusive, diverse, and neutral manner is another reason for the popularity of news engines.

© IEOM Society International

The main benefit of a news search engine is that the engine offers news from multiple sources instead of the regular news websites that offer only the news that they wrote. This allows the consumer to read the news from a different point of view to mitigate the bias that each source may have in its news. Another benefit is that the consumer will be able to see more topics that one source cannot cover them all.

These news search engines are developed based on a set of techniques. They periodically send "web crawlers" to fetch news articles from the news sources, analyse them and then update the news indexes at their servers. A news article is treated as if it was a regular web page posted on the Internet. There is an extensive list of news search engines available currently on the web. Some of the popular ones are: Google News, Yahoo News, NewsBot, and NewsNow.

TAKA (www.taka.media.com) is a new developed news search system based on more than one crawling algorithms. TAKA provides a wide range of functions that including: news aggregation, inquiring, user's interests-based filtering, old news archiving, managing news sources, and reporting. With that comprehensive list of functions, TAKA system can be beneficial for various categories of users: individuals and businesses/organizations. Individuals can use TAKA system to follow up their own interests. Businesses can utilize TAKA to follow up the company image in media's and press's eyes, market changes and opportunities, and as well as their competitors' updates. As any other system, it becomes necessary to measure the quality of the news search systems which can help the systems owners and managers to enhance the performance, achieve the business goals and impress the news-consumers with a high service level.

TAKA is not only a news search engine that aggregates the news in one place. TAKA applies Artificial Intelligence (AI) techniques to analyze the news in many ways. Some of these techniques are topic modeling, classification, morphology, and syntax. Those techniques allow TAKA to present the news to its consumers according to their needs and in the way they prefer it. Moreover, it allows analyzing the whole news to give summaries and aggregated reports to let the consumes understand the trends, topics, size and the quality of the news. This information is the key to measure the influence of the news as well the impact on the people's life. Furthermore, an ongoing research in TAKA is to apply AI techniques for finding fake news and measuring the truth of each article as well. This would play a key role to measure the trustworthy of each resource.

1.2. Related Work

Traditionally few work has been done in the literature to conduct performance evaluation studies for news search engines. However, all those studies focused only on one perspective: the end-user point of view and they ignored the rest of the processes in the news search system.

Originally, Gulli (2005) introduced a general framework to build a news search engine by describing a new academic news search engine. He also presented the components and the architecture of the news search engine. Liu et al. (2007) introduced some features of a news search engine. Furthermore, they reported the results of a comparative evaluation of three commercial news search systems based on several measures: effectiveness, redundancy, diversity, time-sensitivity effectiveness and information richness.

Similarly, to Gulli (2005), Öcalan (2009) presented the architecture, data and file structures, the implementation details, the various features and capabilities, and experimental foundations of the news portal. Moreover, a test collection is developed to enable empirical assessment of new event detection and tracking algorithms. As extension to Öcalan's work, Uyar (2009) tackled the news duplication issue in the news portals as they aggregate news from various sources. He emphasized the existence of duplicate or near-duplicate news in the news portals as a widespread problem, which decreases the efficiency and effectiveness of news search engines. In his master thesis, he proposed and evaluated a new near-duplicate news detection algorithm. In this algorithm, document signatures are created and documents sharing the same signatures are considered as a near-duplicate.

Particularly focusing in ranking of news articles, Curtiss et al. (2009) invented a method for improving the ranking of news articles based on the quality of the news sources. A method for determining a quality of a news Source is provided. The quality indicator associated with each news source is calculated by considering a group of metrics for each news source. Each metric measures a specific characteristic of the news source that could be considered as a partial quality indicator of the news source. The group of metrics was including the number of articles produced by the news source during a given time period, an average length of an article from the news source, the importance of coverage from the

© IEOM Society International

news source, a breaking news score, usage pattern, human opinion, circulation statistics, the size of the staff associated with the news source, the number of news bureaus associated with the news source, the number of original named entities the Source news produces within a cluster of articles, the breath of coverage, international diversity, writing style, and the like.

In order to present an introductory explanation to the news search systems, Doğan (2013) presented in detail the working mechanism, news clustering, news ranking and results extraction in news metasearch engines.

Recently, Bokhari and Adhami (2015) evaluated four news search systems under a new criterion-information richness, i. e., extracting the useful contents from search result record pages and used it for effective evaluation. As extension, Bokhari and Adhami (2016) evaluated the four news search systems using a scheme to find how well the new search systems retrieve fresh news documents. Their scheme used a hybrid criterion considering recall, precision, time-sensitive and relative freshness-based evaluation measures. They compared between the four news search systems based on the top ten results for 100 news queries on the four news search engines.

1.3. Paper Contribution

However, all the previous evaluation studies focused only on one perspective: the end-user point of view. And they ignored the rest of the processes in the news search system. In this paper, we develop a performance evaluation system covering the entire system. Thus, we consider both the end-user's perspective and the entire system perspective as a more holistic view. For this purpose, we develop an ontology-based performance evaluation system for TAKA news search system using the enterprise engineering approach Design and Engineering Methodology for Organizations (DEMO). Henceforth, the structure of the paper is as follow. First, we develop a conceptual DEMO model for the TAKA news search system to describe, understand and analyze the key processes in the TAKA system. Second, we propose a performance evaluation system for the TAKA news search system by developing a set of measures for each process. Third, we analyze the performance and draw potential improvements for the TAKA news search system. Finally, conclusion and recommendations are summarized.

2. DEMO-based System Description

This section provides a detailed description for the core processes in TAKA system. Furthermore, it illustrates the interrelation between the core processes by developing an ontological DEMO construction model.

2.1. TAKA System Description

TAKA is a comprehensive news monitoring system for retrieving, ranking, indexing, classifying and delivering personalized news information extracted from the web. TAKA compiles, filters, abstracts and analyzes news to extract the required news for each user in a customized and efficient way.

TAKA working procedure is as follow: It is continuously working to monitor and aggregate online news. Then it abstracts the news pages to keep the important and essential contents only. It also customizes and filters the news based on user's own interests. This, indeed, makes it easier for individual users and companies to pursue information that most interesting for them. The core processes in TAKA system are presented in more details as follow:

Websites/Sources Database Managing: This process works on accommodating any new websites/sources, fixing and updating or deleting the stopped ones, and making sure that all the websites/sources are working properly. This process aims to enhance the quality of the database and increase the number of news sources as possible.

News Fetching and Aggregation: TAKA gathers news from a set of news sources (Websites/Sources Database) which are updated consciously by the websites/sources database team. Currently, the database contains about 2040 sources. For news aggregation, six algorithms (crawlers) crawl to the websites for fetching the latest news. Crawlers aim to fetch the largest number of news and as fast as possible. A particular algorithm is carefully assigned to crawl for a particular set of news sources based on their historical performance and their functionality. Each of which is systemized to do crawling at particular times during the day. The time between two consecutive crawling is changeable from one algorithm to another, and it depends on each algorithm performance and resources consumptions.

Articles Abstraction: Due to the fact that almost all news links and pages contain unnecessary contents for the user (such as, advertising materials, other news titles, links to other news categories, headers, footers, etc.), abstraction

© IEOM Society International

process aims to extract the main article contents and avoid any other unnecessary materials. Algorithms must extract only required news by increasing readability of the algorithms. This reduces the percentage of spam and inaccurate news.

Articles Analyzing and Filtering: In this process, the aggregated news is analyzed to know the contents types (i.e., text, image, video, …). Furthermore, the article's keywords are identified. Based on the which, an article is classified into one or more of categories (i.e., politics, social, economy, health, woman, technology, sport, …). This process also categorizes the gathered articles based on locations and identifies the interrelated articles. Moreover, the articles are filtered and classified based on users' inquiry's keywords. Thus, this process identifies the news that are relevant to a user, sorted and presented in a convenient way.

2.2. The DEMO Construction Model

In order to develop an efficient performance evaluation system, a clear description and representation of business processes is essential. For this purpose, DEMO is applied for the conceptual modeling and ontology level representation for TAKA system. The Enterprise Ontology DEMO (Dietz 2006) is a well-designed process modeling methodology for enterprise redesign and reengineering, with a special emphasis on enterprise ontology rather than implementation. DEMO is a highly abstracted ontology describing how an enterprise is constructed and enables to create an essential concise view of overall business (Dietz 2006). DEMO abstracts the real world without considering implementation details, so that to grasp the essence of the system. In other words, by defining DEMO model, the core and the stable parts of system can be understood. As such, they present a potentially rich domain for enterprise approach DEMO (Design and Engineering Methodology for Organization) as a modeling approach. The DEMO Construction Model (CM) is the most concise model that gives a description about how transactions and actors' roles are composed to construct a system. In this paper, the CM is developed for TAKA system and it is to be presented next in Figure 1.

The CM is useful for understanding the ontological aspects of TAKA system. The CM is the basic diagram of DEMO that shows the ontological transactions linked to the business roles. Furthermore, it shows the initiator and executor for each transaction. The CM includes four transactions, which represent the core processes in TAKA system: *Articles Analysing*, *Articles Abstracting*, *News URLs Fetching*, and *Websites/Sources Database Managing*. This set of transaction types are abstracted in to {*T1*, *T2*, *T3*, and *T4*}, respectively, and expressed as a disk with a diamond as in the CM shown in Figure 1.

Another main element in the CM is actor role which responds for making decision, commitment or producing products in a transaction. Actor role acts as either initiator or executor of a transaction. There are two types of actor roles, elementary actor role (A) and composite actor role (CA). Elementary actor role is the actor role inside boundary of focus system. In TAKA system case there are four elementary actor roles {*A1*, *A2*, *A3*, and *A4*}. For example, *Articles Analyzer A1*, is the executor of transaction *T1* and the initiator of transaction *T2*. It means that *A1* responds to complete *Articles Analysing* for *customer CA1*, by executing *T1* and initiating *T2*. On the other hand, composite actor role represents a composition of actor roles that are not focused, i.e. *CA1* in the TAKA system case.

Based on the description, we developed the CM of TAKA system using DEMO as shown in Figure 1. The *customer* (*CA1*) is the initiator of the first transaction *T1* (*Articles Analysing*). Because customer is considered as an external actor role, it is shaded (DEMO legend). The executor of *T1* is the *A1 Articles Analyzer*. The executor of any transaction is always differentiated by the red diamond on the link to its transaction. The same actor role *A1* is the initiator for the second transaction *T2 Articles Abstracting*, because the *Articles Analyzer* asks the *Articles Abstracter* to deliver the aggregated articles in an abstracted format in order to be able to analyse them. The *Articles Abstracter* is the executor of *T2* (has the red diamond). In order to complete the *Articles Abstracting* (*T2*), *A2* (*Articles Abstracter*) initiates another transaction *T3* (*News URLs Fetching*). Consequently, the *A2* is the initiator for *T3* (*News URLs Fetching*) and the executors for this transaction is *A3 News URLs Fetcher*. The *Websites/Sources Database Managing* transaction is slightly different. This transaction is almost a routine and continuous process and should be executed by default from the *Websites/Sources Database Manager* in hourly and daily bases, so this transaction is a self-initiated transaction. Where the *Websites/Sources Database Manager* is the initiator and the executor at the same time. In *News URLs Fetching* transaction *T3*, *News URLs Fetcher A3* crawl and visit the set of Websites/Sources Database in TAKA system. Thus, the *News URLs Fetcher* considers information from the *News URLs Fetching* transaction *T3* through an information link (dotted line).

© IEOM Society International

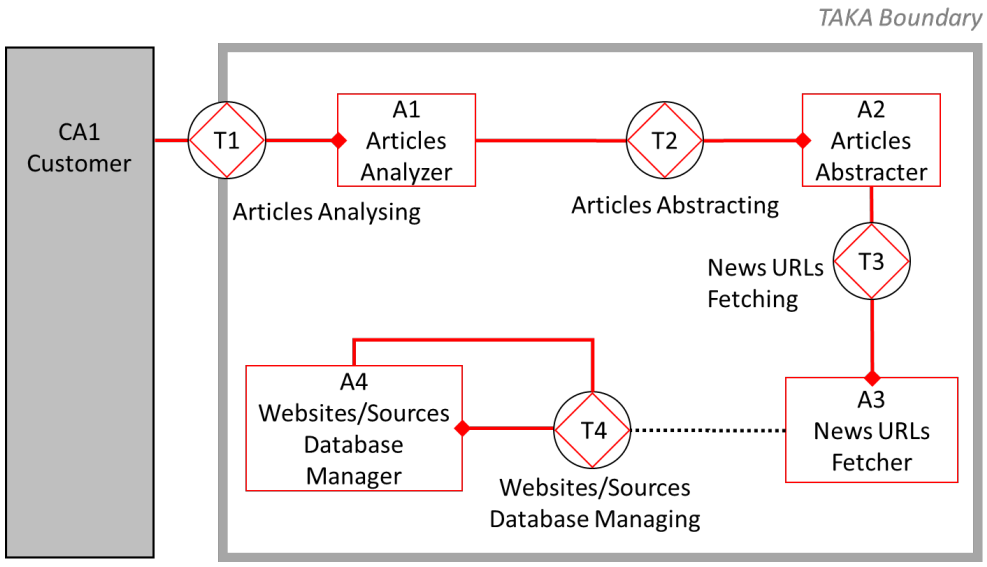

Figure 1. The DEMO Construction Model of TAKA system.

3. The Proposed Performance Evaluation System

In order to develop the proposed performance evaluation system for TAKA, we rely on the DEMO Construction Model as an ontology-based conceptual representation for TAKA system. The proposed performance evaluation system is developed for the key processes in TAKA system based on the DEMO Construction Model. Which reduces the complexity of the performance evaluation system and makes it robust against any changes in the implementation level. It also makes the performance evaluation systems adheres to two main important principles: measuring only what is important and focusing on customer needs.

Referring to the fact that "*what we can't measure, we can't manage*", developing a performance evaluation system is one of the essential needs for any system. This will enable to system control, processes and system assessment, continuous improvement and management assessment. In order to develop that system, we follow a systematic approach with two main steps:

Step 1: Identify the critical processes: First of all, an understanding of what we want to measure is of critical importance. Much effort could be wasted if we do not start with a well-defined system. Then, it is important to focus mainly on the critical processes to be measured. We also need to focus on key areas and processes rather than people. Furthermore, we need to be sure that those key processes are related, directly or indirectly, to the ultimate goal of customer satisfaction. This step is already done in the previous section by developing the DEMO construction model shown in Figure 1, which illustrated the key four processes in TAKA system.

Step 2: Establish Performance Measurements and Metrics: In order to assess the performance for each of the key processes identified in *step 1*, we develop a set of Key Performance Indicators (KPIs). Most importantly, the developed performance measures should reflect results and outcomes of the process, not the activities used to produce results. Furthermore, they should be related directly to a performance goal, based on measurable data, and containing normalized metrics for benchmarking. The proposed KPIs for each of the four critical processes in TAKA system are described in detail below and summarized in Tables 1-4, respectively.

3.1. The proposed KPIs for the Websites/Sources Database Managing Process

As mentioned earlier, this process aims to enhance the quality of the database and increase the number of news sources as possible. Thus, the system must have a fully comprehensive database with a considerable diversity in the sources. The system must monitor and update the news sources periodically to ensure that all the websites are working properly. In order to be able to measure and evaluate the performance of this process, a set of four KPIs for this process is proposed: (1) The total number of Websites/Sources in the database (the database size), (2) The number of inactive

© IEOM Society International

Websites/Sources, (3) The number of active Websites/Sources, and (4) The number of Websites/Sources with no fetched results. A description for each is summarized in Table 1.

3.2. The Proposed KPIs for the News Fetching and Aggregation Process

In this process, a set of crawling algorithms visits periodically the websites for fetching the up-to-date news. As stated before, crawlers aim to fetch the largest number of news, get breaking news, and gather all news that interrelated to the consumer interests. Having a fully comprehensive database with a considerable diversity in the sources is essential for the system, however this is not enough. Efficiency of the crawlers and their ability to fetch news is crucial for TAKA. To be able to measure and evaluate the performance of this process, a set of six KPIs for is proposed: (1) The number of unique articles, (2) The number of unique articles for each crawler, (3) The number of websites/sources with fetched news, (4) The number of undiscovered news, (5) The number of ambiguous links, and (6) The number of redundant links. It is important to mention that the KPI (3) is similar to some of the KPIs for the Websites/Sources Database Managing process. This is because the interrelation between the Websites/ Sources Database Managing process and the News Fetching and Aggregation Process. If a crawler is unable to fetch any news from a website, an issue might be in one of them (the website and/or the crawler itself). Thus, this measure is considered for both the two processes. It is important to say that the KPI (4), the number of undiscovered news, is to be recognized by comparing the automatically fetched news with manual searches. A sample set of manual searches are to be conducted in particular consumer's interests. A description for each is summarized in Table 2.

Table 1. The proposed KPIs for the Websites/Sources Database Managing Process.

Measures	**Description**
1. The database size	This indicator represents the total number of websites/sources in TAKA database. The higher this number the more comprehensive the database. This indicator is checked periodically and compared with the previous records to monitor the progress.
2. Inactive sources	This indicator represents the websites/sources that we deactivated the crawling to them for one reason (i.e., not updated sources, non-official websites, not active websites, difficult to crawl to those websites, …). The lower this number the better the performance of this process.
3. Active sources	This indicator represents the total number of websites/sources in the database minus the number of inactivate websites/sources. The higher this number the higher the potential number of news that could be fetched.
4. Sources with no fetched results	This indicator represents the number of websites/sources that we could not fetch any news from them for a certain period (i.e., a week). This number indicates the existence of issues in websites/sources data, websites/sources updates, and/or crawling algorithms.

Table 2. The proposed KPIs for the News Fetching and Aggregation Process.

Measures	**Description**
1. Unique articles	This indicator represents the number of unique news articles that could be fetched by any of the crawlers for a certain period (i.e., a day, a week). The larger the number of unique articles the more efficient the process.
2. Unique articles for each crawler	This indicator represents the number of unique news articles that could be fetched by each of the crawlers for a certain period (i.e., a day, a week). The larger the number of unique articles for a crawler the more effective the algorithm.
3. Sources with fetched news	This indicator represents the number of websites/sources in the database that the crawlers could fetch news from. The higher this number the higher the potential number of news that could be fetched. This is also gives the potential to cover diverse interests of users.
4. Undiscovered news	This indicator represents the number of news articles that the crawlers could not fetch for a certain period (i.e., a day, a week). This number indicates the existence of issues in the websites/sources and/or in the crawlers.
5. Ambiguous links	This represents the number of news links that are without any contents or without relevant contents. Those links may also contain only advertising contents.
6. Redundant links	This represents the number of news links that are with redundant news and contents. Those links may also show obsolete news and contents.

© IEOM Society International

3.3. The Proposed KPIs for the Articles Abstraction Process

This process aims to extract only the main contents of news articles and avoid any other unnecessary contents for the user. This could eliminate spam and inaccurate news. A set of readability algorithms aim to extract the main and important contents in a news article. The algorithms are uniquely designed to visit URL links and extract only the essential contents (i.e., the title, the date, the main text, images if any, number of words). To be able to measure and evaluate the performance of this process, a set of two KPIs is proposed: (1) The total number and the percentage of articles with readability errors (readability errors), and (2) The number and the percentage of articles with readability errors for each crawler (readability errors for each crawler). It is important to mention that the readability errors could result from errors in the previous processes (i.e., using a wrong URL for news articles, or an error in the article source itself). Thus, we need to check the root causes for each error and take a corrective and sustainable action accordingly. A description for each is summarized in Table 3.

3.4. The Proposed KPIs for the Articles Analyzing Process

This process aims to analyze the aggregated news articles in order to classify them accordingly based on the news categories, locations, interrelated sets of articles, breaking news, and relevant to a particular user's inquiries. This could improve user's satisfaction and saving their time from exploring nonrelevant news. A set of news clustering and topic modelling algorithms are applied to analyze and cluster the news articles accordingly. To be able to measure and evaluate the performance of this process, a set of nine KPIs is proposed: (1) The number and the percentage of articles that clustered in a wrong cluster/category (Wrong clustering/categorizing), (2) The number and the percentage of combined sub-clusters or sub-topics in a cluster/topic (combined clusters/topics), (3) False positive related news articles, (4) False negative related news articles, (5) False positive nearby news articles, (6) False negative nearby news articles, (7) Non-accurate positive nearby news articles, (8) Repeated top news, and (9) Redundant top news. A description for each is summarized in Table 4.

Table 3. The proposed KPIs for the Articles Abstraction Process.

Measures	Description
1. Readability errors	This indicator represents the total number of links that the readability algorithms failed to extract their main contents for a certain period (i.e., a day, a week). The percentage for the number of articles with readability errors is calculated as well. The failure here could be one of the following types of misses: missing the article title, missing part of the main article text, and/or missing images if any. The larger the number and the percentage of articles with readability errors the less efficient the process.
2. Readability errors for each crawler	This indicator represents the number of links that with readability errors for each crawler for a certain period (i.e., a day, a week). The percentage for the number of articles with readability errors for each crawler is calculated as well. The larger the number and the percentage of articles with readability errors for each crawler the less the accuracy of the crawler algorithm.

Table 4. The proposed KPIs for the Articles Analyzing Process.

Measures	Description
1. Wrong clustering/ categorizing	This indicator represents the total number of articles that allocated in a nonrelevant cluster or category for a certain period (i.e., a day, a week). The percentage of this number is also calculated. The larger the number and the percentage the less efficient the clustering and topic modelling algorithms used in this process.
2. Combined clusters/topics	This indicator represents the number of sub-clusters or sub-topics that are combined in one cluster/topic however they could be considered as separate clusters/topics. The percentage of this number is calculated as well.
3. False positive related news	This indicator represents the percentage of articles that the algorithms showed they are interrelated, however they are not actually relevant to each other.
4. False negative related news	This indicator represents the percentage of articles that the algorithms showed they are not related to any other articles, however they are actually relevant to some other articles.
5. False positive nearby news	This indicator represents the percentage of articles that the algorithms showed they are related to a particular area/location, however they are not actually related to that area/location.

© IEOM Society International

6. False negative nearby news	This indicator represents the percentage of articles that the algorithms showed they are not related to a particular area/location, however they are actually related to that area/location.
7. Non-accurate positive nearby news	This indicator represents the percentage of articles that the algorithms showed they are related to a specific area/location. However, they are not actually related to that specific area/location, they are related to a nearby area/location in the same wider area.
8. Repeated top news	This represents the number of news links that are represented/repeated more than once in the top news list.
9. Redundant top news	This represents the number of news links that are redundant/obsolete news and showed in the top news list.

4. Evaluation and Discussions

In this section, we report our evaluation of TAKA news system described in Section 2 using the proposed performance evaluation system introduced in Section 3. Next; the dataset used for this evaluation, the evaluation results and discussions are presented.

4.1. Dataset Used in Evaluation

Data are collected based on the daily and weekly reports extracted from TAKA system. Those reports summarize a wide range of statistics like: the total news fetched, the total sites visited, broken sites, algorithms results and errors, crawling time to gather news for each algorithm, results count by language, and errors percentages. Furthermore, we used data collected based on a set of queries for particular consumers. In order to evaluate the accuracy of the clustering and the interrelationship of and between news, queries results are compared with manually news search by our Articles Analyzing Process team. Around 30 queries or search key words are used in the TAKA automated search and manual search comparisons. The results shown below are based on typical days in October/November 2018.

4.2. Evaluation Results

In this subsection, we report the evaluation results based on the proposed evaluation criterions described in Section 3. Figure 2 shows the number of sites (sources/websites) in TAKA system database. The figure also shows the number of active sites that the crawlers could fetch news from and those inactive/broken sites that the crawlers couldn't fetch any news from. The inactive sites include the sites that we deactivated the crawling to them for one reason (i.e., not updated sources, non-official websites, not active websites, difficult to crawl to those websites, …) and the sites that we could not fetch any news from them for a certain period due to an issue in the sites themselves and/or the crawling algorithms. The figure compares between two different weeks and shows how the database management team could improve the performance of this process. For the first week (the left column for each measure), the total number of sites (sources/websites) in TAKA system database is about 2,040 sites. The crawlers could fetch news from 1,223 sites only, however 818 sites were broken with 60% active sites. The performance is improved for the second week (the right column for each measure) by increasing the active sites to 1,360 with 67% active sites.

Figure 3 shows a simple histogram reflecting the number of weekly aggregated news. TAKA reports show that TAKA system could fetch between 30,000-50,000 articles per day. For a typical week, the figure shows that TAKA system could fetch 378,804 articles in total. However, only 326,415 articles are unique results (many crawlers could fetch the same article) and 19,584 articles are fetched with errors. Those errors could be: links without any contents or without relevant contents; links with only advertising contents; links with redundant/obsolete news and contents; articles with readability errors; missing the article title; articles with errors in the title; missing part of the main article text; and/or missing images if any.

© IEOM Society International

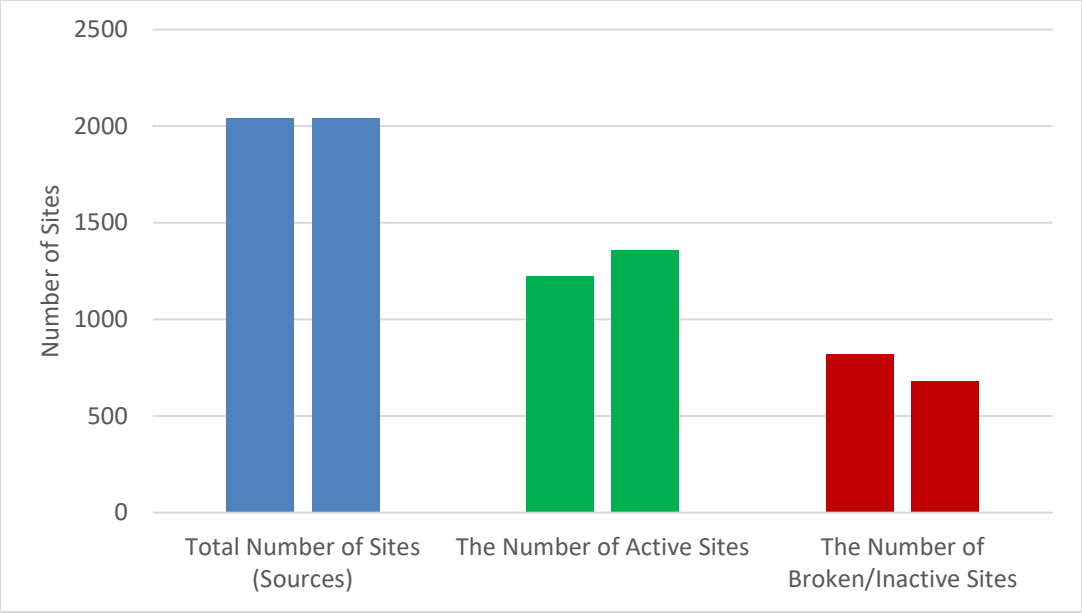

Figure 2. TAKA database statistics summary for two different weeks.

Figure 3. The aggregated weekly results statistics.

Figure 4 shows a pie chart that illustrates the number and the percentage of unique articles for each crawler over a week. The figure compares between the six crawlers used in TAKA system. For a typical week, the results show that the PSA1 and PSA2 come first with more than 30% each of the fetched news. The TWT and the LSA come third and fourth with about 21% and 9% of the fetched news, respectively. Lastly, the SQN and the RSS come last with 4% and 2%, respectively. The percentage errors for each crawler is also reported over a week. However, the TWT contributes greatly in news fetching by around 21%; it comes with the lowest accuracy as it brings around 24% of links with errors. TAKA technical teams are analyzing the reasons behind this low accuracy and the targeting to enhance it.

© IEOM Society International

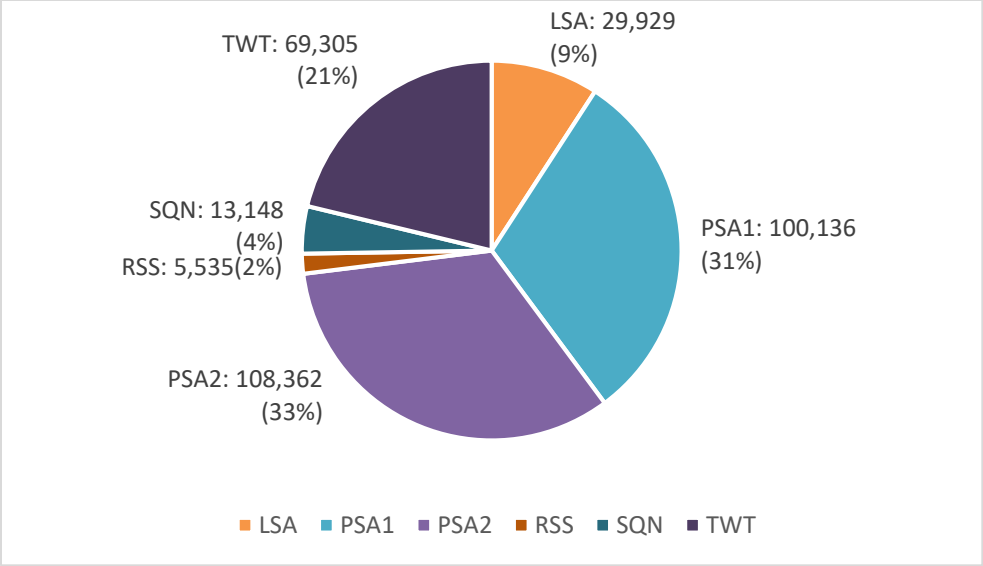

Figure 4. The number and the percentage of unique articles for each crawler.

Figure 5 shows a set of line charts that illustrate the number of news articles by category. The charts illustrate weekly profiles for six news categories extracted from TAKA system. For a typical week, the Politics news is the dominant category with the largest portion of the fetched news along the week. Business and Economy related news come second; Sports news come third; and Health news come fourth. A general pattern could be identified for those four categories that the largest number of news are released on Monday (Day 1) and the number of the daily released news is declining or be steady over the rest of the week. For the last two categories (Environmental and Technology), they are almost time-insensitive and the number of daily news in those two categories are almost steady over the week.

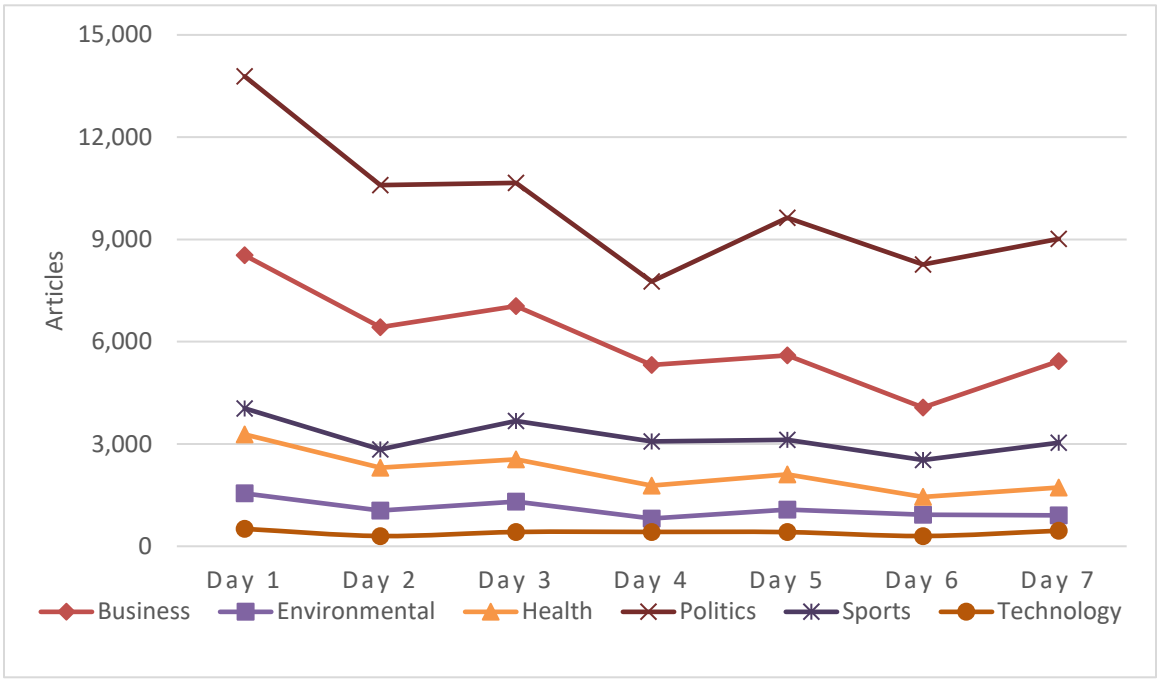

Figure 5. The number of articles by category over a typical week.

© IEOM Society International

Table 5 summarizes the KPIs values for the Articles Analyzing Process. This process aims to analyze the aggregated news articles in order to classify them accordingly based on the news categories (as shown in figure 4), locations, interrelated sets of articles, breaking news, and relevant to a particular user's inquiries. This set of KPIs measures the accuracy of classifying and clustering the fetched news. Results show that the percentage of articles that clustered in a wrong cluster/category is about 30%. The percentage of combined sub-clusters or sub-topics in a cluster/topic is about 16%. This means that the total number of the clustered topics could be increased by 16%. This could be achieved by clustering the sub-clusters that found in a cluster to separate additional clusters. For example, in technology news, a cluster about "New features" is found. However, analysis showed that this cluster could be sub-classified to sub-clusters, where one individual cluster for the new features news for each app or software. Results also show that there are about 30% of news are located to a false or not-accurate nearby location. It is important to mention that the topic modelling and clustering algorithms are still under research and development in TAKA system. Applying the proposed performance evaluation system enabled us to emphasis the shortage in this process performance. However, TAKA technical teams are targeting improving the performance of this process very shortly.

Table 5. The KPIs values for the Articles Analyzing Process.

Measures	**Value**
The percentage of articles that clustered in a wrong cluster/category	30%
The percentage of combined sub-clusters or sub-topics in a cluster/topic	16%
False positive related news articles	19%
False positive nearby news articles	30%
Repeated top news	08%
Redundant top news	17%

4.3. Discussion

Generally, the evaluation results show that TAKA system performs well in most functions. However, during the study, we did observe number of strengths, weaknesses and challenges, which are discussed as below.

Applying the proposed performance evaluation system enabled us to identify and detect the performance shortages in each of the four processes. Furthermore, it enabled the traceability of those issues. Root cause analysis is conducted to investigate the reasons behind those shortages. Then, a set of corrective and proactive actions are taken to handle those issues. TAKA technical teams are targeting to automate the traceability system to be able to detect, analyze, investigate and take actions for any of those issues could exist in TAKA system.

Moreover, applying the proposed performance evaluation system emphasized the challenges of analyzing Arabic texts in comparison with other languages like, i. e., English. This issue is one of our main concerns because we target the Middle East market. The Arabic language, as a natural language, is more challenging in clustering and topic modelling rather than other languages. This could be explained due to the different writing styles and the lack of standards.

It is important to mention that results, in particularly Figure 5, show that media focuses a lot in politics and less about health, environmental and technology. This could raise different research questions about the interrelationship between the media main concerns and the development of our life.

5. Conclusion

In this paper, an ontology-based performance evaluation system for TAKA system is developed. The enterprise engineering approach DEMO is applied to describe, understand and analyze the main processes in TAKA system. Then, a performance evaluation system is proposed for TAKA system by developing a set of KPIs for each process. The paper contribution is to present a performance evaluation system for the entire news monitoring system from user, enterprise and systems engineering point of views. On contrast, studies in the literature consider only the users' perspectives. This study enabled us to identify and detect the performance shortages in each of the four processes. Furthermore, it enabled the traceability of those issues. Moreover, it emphasized the challenges of analyzing Arabic texts. It is also showed that media focuses a lot about politics and less about technology. Future work may focus on tackling those issues. Next, we may analyse the media main focus in each country/region and investigate the interrelationship between the media main concerns and the development.

© IEOM Society International

Acknowledgements
The authors wish to acknowledge the University of Johannesburg for providing financial support. The authors would like also to acknowledge SMAgroup teams and, in particularly, TAKA development and technical teams for the direct technical support and the rich discussions.

References

Bokhari, M.U. and Adhami, M.K., A new criterion for evaluating news search systems, *Communications*, vol. 2, no. 7, pp. 28-35, 2015.

Bokhari, M.U. and Adhami, M.K., How well they retrieve fresh news items: News search engine perspective, *Perspectives in Science*, vol. 8, pp. 469-471, 2016.

Curtiss, M., Bharat, K. and Schmitt, M., Systems and methods for improving the ranking of news articles, U.S. Patent 7,577,655, Google Inc, 2009.

Dietz, J.L., *What is Enterprise Ontology?* Springer Berlin Heidelberg, pp. 7-13, 2006.

Doğan, G., From Newspapers to News Search Systems: Turning Libraries "Inside-Out", *Proceedings of the 21st International BOBCATSSS Conference*, Ankara, Turkey, January 23-25, 2013.

Gulli, A., The anatomy of a news search engine, *In Special interest tracks and posters of the 14th ACM international conference on World Wide Web*, Chiba, Japan, May 10-14, 2005.

Liu, K.L., Meng, W., Qiu, J., Yu, C., Raghavan, V., Wu, Z., Lu, Y., He, H. and Zhao, H., AllInOneNews: development and evaluation of a large-scale news metasearch engine, *In Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, Beijing, China, June 11-14, 2007.

Öcalan, H.Ç., *Bilkent News Portal: A system with new event detection and tracking capabilities*, master's thesis, Computer Engineering Department, Bilkent University, 2009.

Uyar, E., *Near-duplicate news detection using named entities*, master's thesis, Department of Computer Engineering, Bilkent University, 2009.

Biographies

Tarek Fatyani is a Researcher and a Business Analyst. He received his PhD in Industrial Engineering and Management from Tokyo Institute of Technology 2016. His research was about business process analysis and how is it related to software development. He established his company in Japan SMAGROUP LLC in 2015 while he was a PhD candidate. His research and experience are related to virtual collaboration tools, innovation management, project management, enterprise engineering, modelling and software development. Currently, he is focusing on utilising machine learning and data analysis to advance the software that he is leading.

Zakaria Yahia received the M.Sc. and Ph.D. degrees in Industrial Engineering from Cairo University, Giza, Egypt, in 2012 and Egypt-Japan University of Science and Technology (E-JUST), Alexandria, Egypt, in 2015, respectively. As a visiting Ph.D. student, he spent one academic year at the Tokyo Institute of Technology (TITECH), Tokyo, Japan, working on the research project "Developing a Design and Engineering Methodology for Organization (DEMO)-based simulation model for surgery room system". From 2015 to 2017, he was an Assistant Professor with the Department of Mechanical Engineering, Fayoum University, Fayoum, Egypt. Currently, he is a Post-Doctoral Researcher with the Department of Quality and Operations Management, University of Johannesburg, South Africa. His research interests include the areas of Applied Operations Research & Simulation, Scheduling, Healthcare Management, Smart Grid Management and DEMO-Enterprise Ontology.

© IEOM Society International