This article can be cited as K. El Handri and A. Idrissi, A Top_{kws} Algorithm for Synthetics and Real Datasets, International Journal of Artificial Intelligence, vol. 19, no. 1, pp. 36-55, 2021.

A Top_{kws} Algorithm for Synthetics and Real Datasets

Kaoutar El handri [1] **and Abdellah Idrissi**[2]

Intelligent Processing Systems Team, Computer Science Laboratory (LRI)
Computer Science Department, Faculty of Sciences, Mohammed V University in Rabat
kaoutar.elhandri@um5s.net.ma [1], abdellah.idrissi@um5.ac.ma [2]

ABSTRACT

Recommendation Systems (RS) is the most commonly used Information technology in the last decade. RS's processing behavior can be found on different approaches, such as Decision Making (DM) support and preferences query processing. These systems have been used in many Internet activities, mainly to overcome information overload and for many other purposes. Some of these include e-commerce sites, web page searching, e-learning, and Cloud Computing Services. Also, research has been conducted on the use of RS in some sport management Systems. This paper presents a performance evaluation of the $Topk_{ws}$ recommendation algorithm, applying on a new RS called Generic Research and Selection System (GRSS) based on the Skyline and the Top_k query processing. The conceived algorithm, which used an adapted Multi-criteria Decision Aiding (MCDA) method, was applied to different research eras in this paper, namely the Cloud Computing Services and Sport management System. The algorithm shows to be more and more efficient. Extensive experiments based on correlation study toward both real and synthetic datasets demonstrate the efficiency and scalability of this algorithm compared with other best-known algorithms in this field.

Keywords: Recommendation system, Multi-criteria Decision Aiding (MCDA), $Top_{k_{WS}}$, Generic Research and Selection System (GRSS), Skyline.

2012 Computing Classification System:

- Human-centered computing $\longrightarrow$ Human computer interaction (HCI)
- Information systems $\longrightarrow$ Information retrieval
- Computing methodologies $\longrightarrow$ Artificial intelligence

1 Introduction

In the last decade, various works proposed their RSs for Top_k query processing. An important issue in such systems is faced by the Big Data size and dimensionality, with a sufficient level of quality. The principal purpose of RS is to help each user identify the ideal results of a manageable size. Let consider an RS based on each user's most required videos during the football World Cup event (FIFA, 2018 (accessed August 03, 2018)) (Wordcup, 2018 (accessed April 15, 2020)). The given statistics show that some users needed to follow the match between the finalist, while others worried about seeing a movie or a cartoon video, accordingly. It depends on the user's interest and feedback. The users find their requirements

acting on their systems. Therefore, a subset of the most relevant answers instead of all solutions is given using the recommendation query. Not far from this example, and to provide an accurate ranking that is explicitly applied to the impact of the 2018 FIFA World Cup matches (Burer, 2012; Saeed, Saqlain and Riaz, 2019). RS's Man-of-the-match prediction is also a fascinating and seductive research problem, not least because of its complications, the effort it requires, and unexpected results. Hence, a football match depends on various factors, characters, and irregular situations. As a result, it is difficult to predict the correct and accurate results of football matches. Such research requires a Multi-Criteria Decision-Making (MCDM) approach to predict a ranking that is explicitly applied to the 2018 FIFA World Cup football matches' outcomes. The application of the GRSS has proven that correlations results between the different criteria are significantly better than those of the other algorithms. However, from state of the art, the Top_k (Idrissi, El handri, Rehioui and Abourezq, 2016), and the Skyline (Idrissi and Abourezq, 2014) are considered the most popular used questions in the information retrieval and query preference field. This paper presents an extensive work about $Top_{k_{WS}}$ algorithm based on multi-criteria decision support. The algorithm was being used by applying the Cloud Service Research and Selection Section (CSRSS) (Abourezq and Idrissi, 2014a) shown in figure 1, and that was being applied on the Cloud Computing Quality of Service (CCQoS) (Abourezq and Idrissi, 2014b) dataset presented in Table 1. Then we compare the proposed algorithm with the Threshold Algorithm (TA) and the No Random-access Algorithm (NRA) (Fagin, Lotem and Naor, 2003), (Fagin, Kumar and Sivakumar, 2003). The performance evaluation of the studied algorithm, which is the $Top_{k_{WS}}$, is performed by using two subsystems of recommendation in both synthetics and real datasets. The subsystems above, namely the Recommender System number 1 (RS1) using the CCQoS dataset and Recommender System number 2 (RS2) using the FIFA dataset, are the component of a comprehensive System named the Generic Research and Selection System (GRSS), which is an improved approach of the CSRSS presented in (Abourezq and Idrissi, 2014a; El handri and Idrissi, 2020a). This paper is structured as follows: In section 2, we expose the related works. In section 3, we present our approach using two subsystems of recommendation. In Section 4, we extend experimental results based on real and synthetics datasets. Section 5 presents a brief discussion of the result achieved during these experiments. Then, the conclusion and some future work are given in section 6.

2 Related Works

2.1 Background and related works

Collaborative filtering and query processing methods have expressed significant evolution from the past decade. These tools are used in recommendation systems that require innovative platforms (Ott, Hayashi and Fukuda, 2007) to incorporate the user behavior (Vaziri, Dabadghao, Yih and Morin, 2018). Besides, easy access to information and descriptions of desired services (e.g., restaurants, hotels, department stores, and travel destinations) has led to many decision support systems and recommendation services. The most used algorithms for query processing like Top_k recommendation algorithm for refining the Skyline (Liu, Xiong, Pei, Luo,

Zhang and Yu, 2019) query for multi-dimensional data can lead either to a very large or to an insufficient number of responses. Indeed, these results can confuse the choice of user (Idrissi et al., 2016). Therefore, before discussing the related works that use the Skyline and the Top_k algorithm in decision aiding and RSs, let us shed light on the objective of using this paradigm. Meanwhile, the Skyline, also referred to as Maximum in computational geometry or Pareto in business management, is essential for many applications where multi-criteria decision-making is needed.

Let have P a set of points n. Each point p of d real-valued attributes can be described as a point $(p[1], p[2], \ldots, p[d]) \in \mathbb{R}^d$ where $p[i]$ is the ith attribute of p.

Now, let's assume that $p = (p[1], p[2], \ldots, p[d])$ and $p' = (p'[1], p'[2], \ldots, p'[d]) \in \mathbb{R}^d$, p dominates p′ if for each i, $p[i] \leq p'[i]$ and for at least one i, $p[i] < p'[i] (1 \leq i \leq d)$.

The Skyline is determined as the set of points P that are not dominated by another point of P (Kalavagattu, Das, Kothapalli and Srinathan, 2011). Meanwhile, the Skyline performs the most important points or the optimal Pareto solutions of the data set. These Skyline points cannot dominate each other (Liu, Xiong, Pei, Luo and Zhang, 2015). Nevertheless, some first works have been reported in (Tiakas, Valkanas, Papadopoulos, Manolopoulos and Gunopulos, 2016) the principle of dominating query in a dynamic context. A generalization was based on this approach of dominating queries concept into multi-dimensional datasets.

In other studies, authors choose to combine the Skyline and Top_k paradigms in distributed decision Systems. For example, the study in (Amagata, Hara and Onizuka, 2018) presents a decentralized approach that conceived the Space-Filling Algorithm (SFA) for dominating query processing. This algorithm is based on a traditional weighted sum score function. However, it can provide an approximate set of responses and improve query processing efficiency despite improving this algorithm. The algorithm enhancement was attained by sacrificing accuracy performance.

In another context, handling Top_k problem was considered the several employed in this domain, such as Fagin's algorithm's amelioration. Wherever the studies in (Fagin, Lotem and Naor, 2003) show that The Threshold Algorithm (TA) is like the Fagin Algorithm (FA) with some improvements, in this algorithm, an approximation function is used to find the optimal degree in all cases. It uses less buff space to stop earlier. Algorithm 1 presents the TA as shown in (Fagin, Lotem and Naor, 2003). Furthermore, another enhancement of FA is given by the No Random Access algorithm (NRA).

Moreover, this algorithm is used for dealing with the situation where the Random Access (RA) is very expensive than the Sorted Access (SA) or is impossible (Fagin, Lotem and Naor, 2003) as it can be seen in algorithm 2.

In general, all algorithms proposed so far should use one type of access among these four categories:

- Random and sequential access: In this category, sources are accessible via random and sequential accesses. Among the proposed algorithms, the best known is the Threshold Algorithm (TA) (Fagin, Lotem and Naor, 2003), which considers access to all sources in turn.

- Without random access. We consider in this category sources accessible only by se-

Algorithm 1 TA

Do parallel SA on all m lists object x seen under SA in a list
fetch its scores from other lists by RA
compute overall score
if $|Buffer| <$ k **then**
    add x to $Buffer$
elseif score(x) $\leq k$ th score in the $Buffer$
    replace bottom of buffer with (x, $score(x)$)
    toss
    Stop when threshold $\leq k$ th score in the $Buffer$
    Threshold := t (worst score seen on L_1,...,L_m)
end if

Output the Top_k objects & scores (in $Buffer$)

Algorithm 2 NRA

Do SA on all lists in parallel
Stop when there are at least k objects, each of which have been seen in all the lists
while depth d **do**
    Maintain worst scores x_1, ..., x_m,
    x any object seen in lists $1,...,i$
    Best(x) = t(x_1, ..., x_i, x_i+1, ..., x_1)
    Worst(x) = t(x_1, ..., x_i, 0, ..., 0)
    Top_k contains k objects with max worst scores at depth d
    Break ties using Best. $M = k-th$ worst score in Top_k
    Object y is viable if $Best(y) > M$
    Stop when Top_k contains $\geq$ k distinct objects and no object outside Top_k is viable
end while
Return Top_k

quential access. The NRA algorithm (Fagin, Kumar and Sivakumar, 2003) is the best known in this category. Like TA, it accesses all sources in turn.

- Sequential access with controlled random access: This third category of Top_k query processing techniques generally considers a source with subsequent access to discover objects. Several unsorted sources were queried by random accesses to calculate the final scores of the items found in the first source. The cost of random access is often considered much more important than sequential access; for this reason, different parameters are used to control and limit random access. This type of access is the case of our conceived Top_{kws} algorithm 3 in which we use score access controlled by a priority queue as Will be discussed after.

- All types of access: The last case is the one that has been the least treated in the

literature. This category represents the generic case, with algorithms handle the different configuration of source types and access costs. An example of the work done in this category is the Necessary Choices algorithm (Hwang and Chang, 2007).

2.2 Previous works in different application domains

To the best of our knowledge, our method which is used to combine both paradigms, namely, Top_k and Skyline, is the first work that uses a Bi-objective Weighted Sum (BWS) to handle the Skyline and Top_k dominating query problem. In the previous work, we suggested applying the adapted weighted sum method as a powerful core of the simple approach of similarity that we will introduce in the following section. The principal objective of this method is to reflect the preference of the end-user accurately. For this reason, we have developed in (Idrissi et al., 2016; Abourezq and Idrissi, 2014a), a system based on the principle of the Skyline. We then tried to improve our policy by applying outranking methods (Abourezq and Idrissi, 2014b) to the results returned by the Skyline. And finally, we use the Top_k query to select an object based on its importance. It's about finding the most relevant to the least concerned Top_k object. The combination of Top_k and Skyline was used in synthetic dataset (El handri and Idrissi, 2020a). The Fagin algorithm's comparison is based on correlation study, and runtime measurement of this approach shows its performance in the Cloud Computing Service domain. Consequently, the presented work aims to evaluate the used algorithm in a general context using the GRSS in the other application domain and using FIFA 2018 as a real dataset. However, sport management needs to take into account different properties (Vaziri et al., 2018). Therefore we adopted the conceived algorithm while using users' choice scenarios for responding top Man of the match according to these characteristics. In practice, match statistics are never available before the match, which leads to using previous match statistics for accurate predictions. According to the authors in (Wheatcroft, 2020), this approach reflects that informative match prediction can be made. The authors used the predicted statistics, which was calculated using a comprehensive method called: The Generalized Attacking Performance (GAP) Ratings. Another sport management research was based on the Generalized Fuzzy Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) as an MCDM tool. The method TOPSIS was presented by Hwang and Yoon in 1981 (Tzeng and Huang, 2011; Lai, Liu and Hwang, 1994). However, Fuzzy TOPSIS, used in the discussed work, was introduced to explicitly predict final ranking and apply it to the results of FIFA 2018 world cup datasets. The match statistics have been used up to quarter-finals to make better estimates for the upcoming games. Our work remains similar to the work mentioned above due to using the most widely used MCDA methods, particularly in sport management. Our approach differs, however, in that we use recommendation algorithms such as Top_k and Skyline to take advantage of their combination and hybridization with MCDA methods in two different application areas.
Nevertheless, apart from the critical interest in fuzzy modulation in decision support (Gil, Johanyák and Kovács, 2018), (Precup and David, 2019), the modeling of meta-heuristic optimization also considers system dynamics. For example, Marung et al. (Marung, Theera-Umpon and Auephanwiriyakul, 2016) use a genetic algorithm to model the problem of sparsity in recommender systems. These main meta-heuristic methods are used to manage dynamic

systems such as work (Purcaru, Precup, Iercan, Fedorovici, David and Dragan, 2013; Osaba, Del Ser, Sadollah, Bilbao and Camacho, 2018). In our work presented in (El Handri and Idrissi, 2020b), we have also addressed this problem by using collaborative filtering methods based on singular value decomposition (SVD) (El Handri and Idrissi, 2020b) as a machine learning technique. However, in this work, we focus on using MCDA optimization with Top_k query processing. However, in our previous works, as illustrated in the introduction, we have chosen to use the ELECTRE IS with Skyline in the first step, then we have used Top_k in two different ways as shown in Figure 1. For more details on the Skyline algorithm's combination, namely the Block-Nested Loops algorithm (BNL) of Skyline combined with the ELECTRE IS algorithm, see papers (El handri and Idrissi, 2019; El handri and Idrissi, 2020a; Abourezq and Idrissi, 2014a; Idrissi et al., 2016).

3 Our approach using the Generic Research and Selection System (GRSS)

The combined approach of Top_k and Skyline processing manages the Skyline drawbacks. Still, the processing step shown in figure 1 cannot always be a suitable solution because it did not cover all possible recommendation scenarios. However, there are still a few scenarios in which the user can determine his / her ranking function, where the Skyline step is not required. The preprocessing step can increase the recommendation runtime. Consequently, this remark drives us to study the presented approach in the general case and to provide a complete extended system solution that the requirements needed in these paradigms.

The GRSS, considered in figure 2, consists of the design of two subsystems that operate

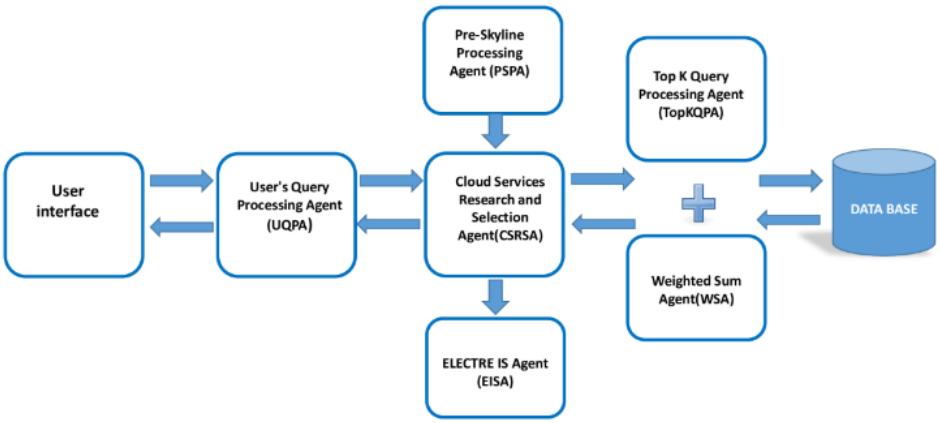


Figure 1: Top_k and WS Agent using for Skyline refining problem in (CSRSS), extracted from (El handri and Idrissi, 2020a)

respectively in two stages. In the following, we give an illustration of formalism.

3.1 System for refinement of Skyline dominating query

In the first stage, the combined approach uses the advantages of both paradigms. The aim is to handle the problems coming from Top_k used over a Big Data size, which creates an expensive sort of Top_k query, and the problem coming from the Skyline on high dimensionality.

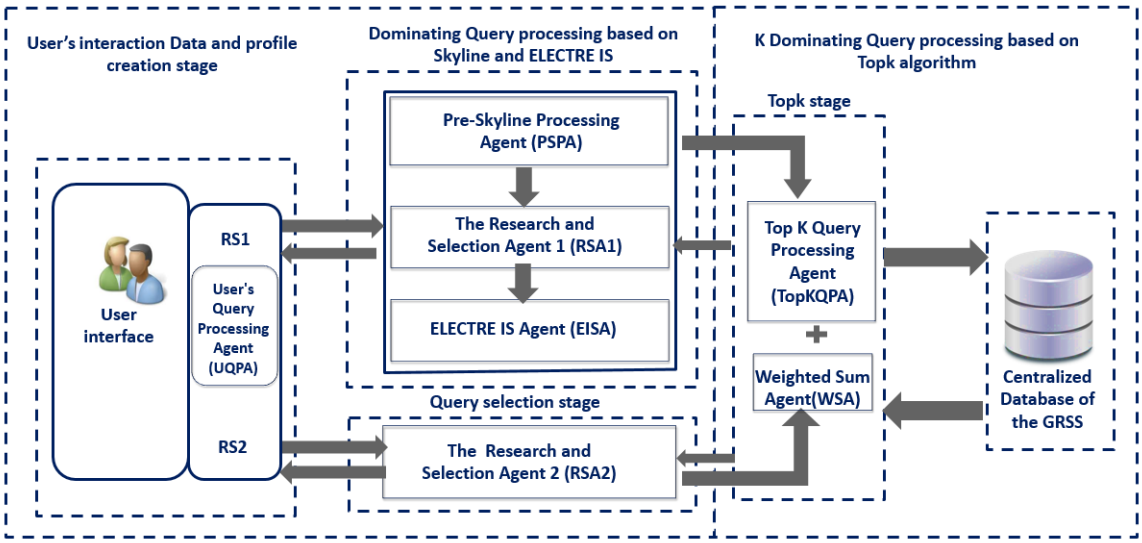


Figure 2: Generic Research and Selection System (GRSS) based on Skyline and Top_k

The Pre-Skyline Processing Agent (PSPA) provides the results from the database for the CSRSA and manages the Skyline Operator. Finally, the best requirements Cloud Services are returned. This stage was presented as a collaborative filtering system by using the Skyline agent, the ELECTRE IS agent (El handri and Idrissi, 2019; El handri and Idrissi, 2020a), and Top_k agent to better optimize the result size and response time of the request, (Idrissi et al., 2016; El handri and Idrissi, 2019; El handri and Idrissi, 2020a) while maintaining efficiency and accuracy of the obtained results.

3.2 System for Top_k dominating Query

In the second stage, we use the $Top_{k_{WS}}$ algorithm directly on the datasets to show the effectiveness of this algorithm, especially when the Skyline is not required.
The power of the applied query processing was incorporated into multi-objective behavior based on an adapted Multi-Criteria Decision-Aiding approach.
Finally, the processing of the GRSS uses the standard treatment of Users Query Processing Agent (UQPA), which uses either the Top_k alone or combine the Skyline and the Top_k query together for the Research and Selection Agent. In the following, we explain the Top_k and the Weighted Sum Method Agents in detail, and we present the $Top_{k_{WS}}$ algorithm.

3.3 Bi-objective Weighted sum Method (BWSM) (El handri and Idrissi, 2020a)

Let R^n and R^p be Euclidean vector spaces referred to as the decision space and the objective space. We denote $X \in R^n$ as a feasible set and f as a vector-valued objective function.
$f : R^n \longrightarrow R^p$ composed of p real-valued objective function $f = (f_1, \ldots, f_p)$, where f_k: $R^n \longrightarrow R^p$ for $k \in \{1, \ldots, p\}$ A multi-objective program (MOP) is given by:

$$min(f_1(x), ..., f_p(x)), x \in X \tag{3.1}$$

The function 3.2 shows the minimization of the weighted sum objective functions:

$$min \sum_{k=1}^{p} \omega_k f_k(x), x \in X, and\ \omega \in R^+. \tag{3.2}$$

Where the weights $\omega_i, i \in \{1,...,p\}$ corresponding to objective functions satisfy the following conditions: $\sum_{i=1}^{p} \omega_i = 1, \omega_i \geqslant 0, i \in \{1,\ldots,p\}$.
we assume the Bi-objective Weighted Sum as BWS to designate the method presented in (Idrissi et al., 2016) and (El handri and Idrissi, 2020a) instead of WS, which is used as a linear weighted sum aggregation function. Consequently, we consider the (MOP) presented in function (3.1) with linear objective functions, having the following form:

$$f_i(x) = \sum_{k=1}^{p} a_{ki} x_{ki}, a_{ki} \in R^+. \tag{3.3}$$

$$R(a_i) = \sum (w_j a_{ij}) \forall i \in [1,n] \tag{3.4}$$

With w_j represents the weights chosen by the user, and the a_{ij} represents the values of the criteria to be maximized or minimized. Therefore, If the criteria are to maximize, the algorithm seeks the maximum at using function eqEM2. If not, as can be seen in function 3.5, the 1 considers the inverse of these values.

$$\begin{cases} max \sum (w_j a_{ij}) & \text{if } a_{ij} \text{ to be maximized} \\ max \sum (w_j 1/a_{ij}) & \text{if } a_{ij} \text{ to be minimized} \end{cases} \tag{3.5}$$

Function 3.6 chows the BWS, which was used to adapt the general WS and represented the aggregation function. This monotonic and linear function allows us to answer a bi-objective problem while optimizing the research cost by providing guarantees on the evolution of the candidates' scores (El handri and Idrissi, 2020a).

$$f : max \sum (w_j a_{ij}) + max \sum (w_j (1/a_{ij})) \tag{3.6}$$

The $Top_{k_{WS}}$ uses the monotonic ranking score. Based on the aggregation function, which is denoted by the scorefunction in algorithm 3. It starts by reading an input tuple from a priority queue. Then it compares it to the Skyline and ELECTRE IS output list LS using the function Compare (Ls, item). The WSM agent is used for computing the score function, as showed in equation 3.6.
Concerning the stage of the Top_k agent, it is based on the Top_k queries that can be modeled as follows:
We denote m a set of n lists of data items such as each data item has a local score in each listing, and the records are sorted according to their local scores. And each data item has an overall rating, which is calculated based on its Local scores in all lists using a given scoring function. Then the problem is finding the Top_k items whose overall scores are the highest according to the formulas given in equation 3.5 and 3.6. If the condition is verified, this means that criteria are sometimes to maximize and sometimes to minimize. These criteria characteristics represent the performance of each action on each of the criteria. While the algorithm parameters are defined as follow:

- we define the criteria as $\mathrm{C1}, \mathrm{C2}, \mathrm{C3} \ldots \mathrm{Cn}$.
- Vector weight $(\omega_1, \omega_2, ..., \omega_n)$ and $\omega_i > 0$.
- a_{ij}=u_i(Ai), cardinal utility function quotient. Based on these starting parameters, we then use an iteration for computing the score function, as mentioned in 3.6.

Thus, at the end of all iterations, the Top_k agent keeps only the best compromise items in all the requirements defined by the user using Top_{kws} algorithm which stops when the Top_k list is computed.
In follow, we will present description of the algorithm's implementation and its performance, as well as some illustration of its application. To assess the evaluation of our approach, we introduced the obtained results in the following section.

Algorithm 3 $Top_{k_{WS}}$

Ls: input PriorityQueue
$tlArray$: array of items
$Item$: input object which will calculate its score function
k: the number of objects returned by the algorithm
$TopList$: output list of the tuples forming the solution
Define Ls as priority queue based on $ScoreFunction$
function $ComputeTopK$
$returned$ =0
while $returned < k$ **do**
    $Ls \neq \emptyset$ $Ls \in PriorityQueue$
    $result = Compare(Ls, item)$
    if $result < 0$ **then**
        Select from Ls the object item with the maximum $ScoreFunction$ using equation (3.6)
        Remove the head of this queue or returns null if this queue is empty
        $Ls.add(item)$
        Update $ScoreFunction(item)$, and update Ls accordingly
    else
$ScoreFunction(item)$ is completely known
        $Report(item, ScoreFunction(item))$ and
        $returned = returned+1$
    end if

end while
return $TopList$
end function

To evaluate the performance of requests, we extract from the state-of-the-art, the most used metrics among this field's ranking methods. We use the correlation study based on Kendall, Spearman coefficients, and their correlation significance coefficients, respectively, based on the p-value. The similarity between two users is based on their ratings of items that both users

have rated (Adomavicius and Tuzhilin, 2005).
Let $[n] = 1 \ldots n$ be a universe of elements. Let S_n be the set of permutations on $[n]$ and for $\sigma \in S_n$, let $\sigma_{(i)}$ denotes the rank of the element i. We have two well-known metrics that evaluate the distance between two permutations $\sigma, \tau \in S_n$: The Spearman's footrule distance $F_{(\sigma,\tau)}$, and the Kendall's tau $K(\sigma,\tau)$. The both coefficient can be represented by Daniels formula as shown in (Saporta, 2006). We consider for every pair of individuals i, j two indices's a_{ij} and b_{ij}, the first one associated with the variable X. The second one associated with the variable Y, for example, $a_{ij} = X_i - X_j$, and n is the sample size. In following we define the coefficients:
The Daniels formula (Saporta, 2006):

$$\frac{\sum_{i,j=1}^{n} a_{ij} b_{ij}}{\sqrt{\sum_{i,j=1}^{n} a_{ij}^2} \sqrt{\sum_{i,j=1}^{n} b_{ij}^2}} \tag{3.7}$$

Definition 3.1. The Spearman's coefficient (Saporta, 2006):
Taking $a_{ij} = r_i - r_j$ and $b_{ij} = s_i - s_j$. We obtain The Spearman's coefficient according to the formula number 3.7 where r and s are the rankings according to X and Y,

$$\rho = \frac{\sum_{i,j=1}^{n} (r_i - r_j)(s_i - s_j)}{\sqrt{\sum_{i,j=1}^{n} (r_i - r_j)^2} \sqrt{\sum_{i,j=1}^{n} (s_i - s_j)^2}} \tag{3.8}$$

Definition 3.2. The Kendall's coefficient: (Saporta, 2006):
Taking a_{ij} = sign of $x_i - x_j$ and b_{ij} = sign of $y_i - y_j$.
With sign of $x_i - x_j$= $\frac{x_i - x_j}{|x_i - x_j|}$ and sign of $y_i - y_j$= $\frac{y_i - y_j}{|y_i - y_j|}$, Kendall's coefficient:

$$\tau = \frac{\sum_{i,j=1}^{n} \frac{x_i - x_j}{|x_i - x_j|} \frac{y_i - y_j}{|y_i - y_j|}}{\sqrt{\sum_{i,j=1}^{n} \frac{x_i - x_j}{|x_i - x_j|}^2} \sqrt{\sum_{i,j=1}^{n} \frac{y_i - y_j}{|y_i - y_j|}^2}} \tag{3.9}$$

4 Experiment results

We use a Core i5 (2.70 GHz) PC with 8 GB of memory in the experiment. Algorithms are implemented in the JAVA environment.

Table 1: The used configuration for the experiments.

dataset Num	dataset Type	Size (items)	Dimension
dataset1	Synthetic (CS QoS)	50000	4 - 10
dataset2	Real (FIFA Predict)	128	4 - 10

The used comparison takes into account the algorithms discussed above, namely, TA, NRA, and Top_{kws}. However, before showing the application of these algorithms and the comparative study, we aim to sum up the principle behind the optimization problem and its relationship with each algorithm, as is already showed in the pseudo-code.

- For the algorithm TA 1, at each sequential access, the algorithm first set the threshold t to be the aggregate of the scores seen in this access. Then it does random accesses and computes the scores of the observed objects. Maintain a list of Top_k objects seen so far. Finally, the algorithm stops when the scores of the Top_k are greater or equal to the threshold. Then it returns the Top_k seen so far.

- For the NRA algorithm 2, it Accesses all lists sequentially in parallel and is stop until there are k objects for which the lower bound is higher than the upper bound of all other objects. Then it returns Top_k objects for which the lower bound is higher than the upper bound of all other objects. The Top_{kws} algorithm 3 uses a priority queue for accessing all list sequentially until there are k objects. Then it sets the record accordingly based on the priority heap. The priority queue elements are ordered by a comparator provided at queue construction time, depending on the given weighting. After that, the algorithm sets the list accordingly and compute the score of the seen objects. The algorithm stops when the score of Top_k is greater than to previously founded score (using a weighted aggregation monotonic function for calculating the rating). Finally, the algorithm returns a Top_k object that satisfied the maximum score and the given weighting.

4.1 Experiment results based on Synthetic datasets

The approach was used on the Cloud Service QoS (CS QoS), which designs the dataset1 in Table 1 using algorithm 3. The denoted algorithm was compared with the Fagin algorithm (FA) in the study presented in (El handri and Idrissi, 2020a) using the same dataset. The comparison showed the efficiency of our approach according to data size and dimensionality variation, which was approved by runtime measurement and six recommendation metrics between rankings. Furthermore, an extended evaluation is given on this experiment based on the comparison with other algorithms, namely the TA and NRA presented above, using the dataset1 and dataset2 shown in Table 1. The studies (Idrissi et al., 2016) and (El handri and Idrissi, 2020a) contains more details about the dataset1.

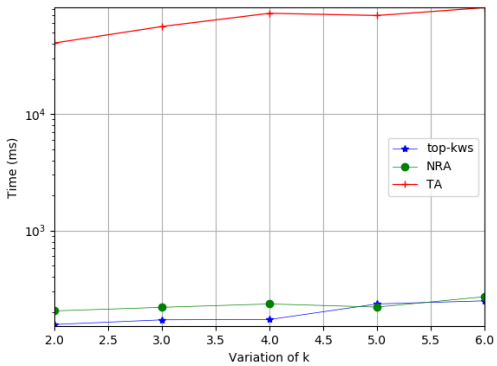
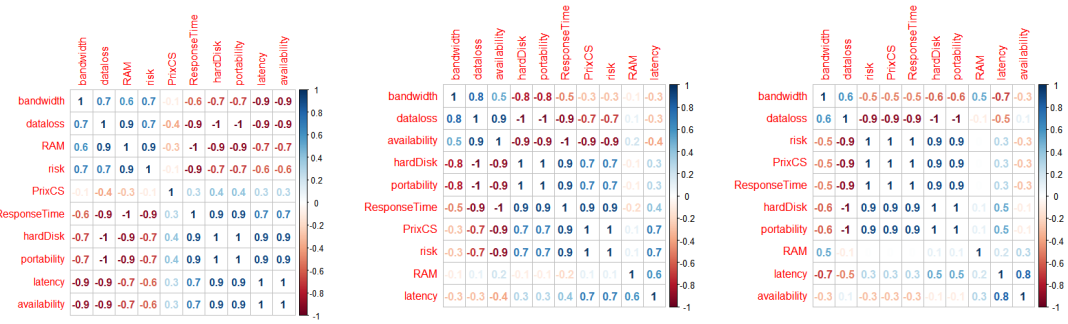
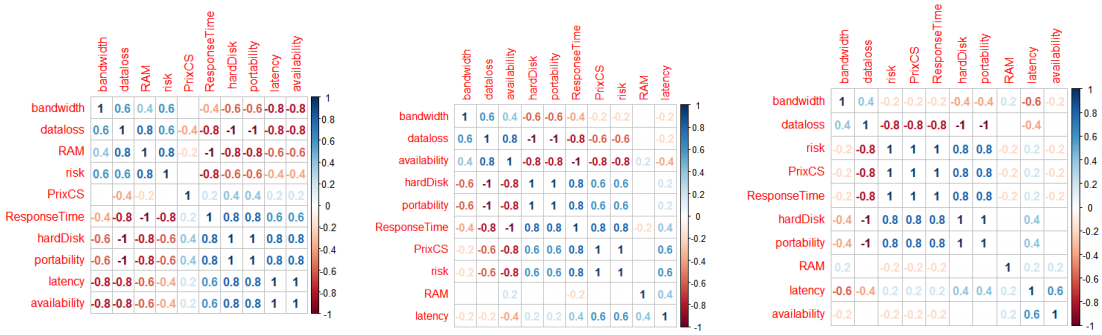


Figure 3: Response time variation of the $Topk_{ws}$, TA and NRA algorithm according to k variation using synthetic dataset



(a) $Topk_{ws}$ (b) TA (c) NRA

Figure 4: Spearman correlation coefficient



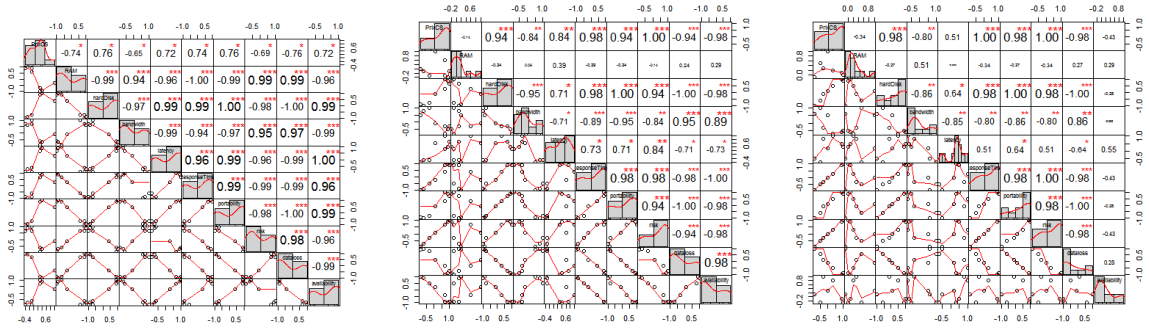
(a) $topk_{ws}$ (b) TA (c) NRA

Figure 5: Kendall correlation coefficient

4.2 Experiment results based on real-life datasets

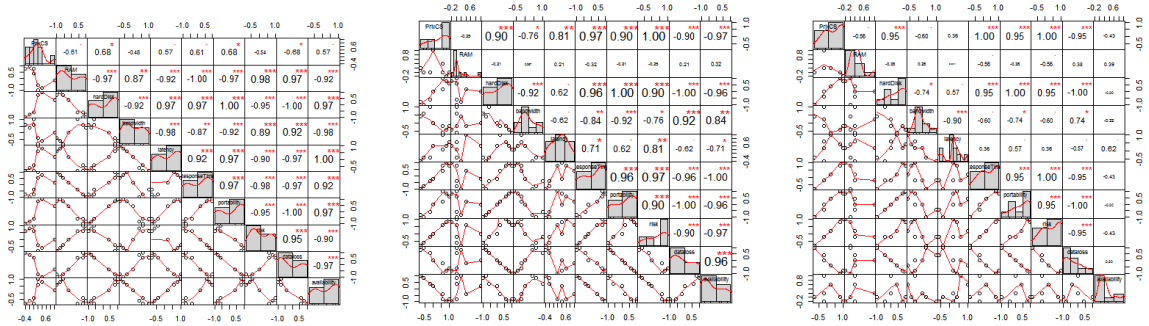
For dataset2, we use the FIFA Predict 2018 Man of the Match found on 3 (FIFA, 2018 (accessed August 03, 2018)) shows an extracted example from the used dataset. It should be noticed that each row contains a team's performance details. For example, a match between Morocco and Iran on 14th June 2018 has two rows: 'Morocco' and another one for 'Iran'.

1. Goal Scored (GS): Number of goals scored by this team
2. Ball Possession % (Percentage) (BallP): Amount of time ball was in control by the team
3. Attempts: Number of attempts to score a goal
4. On-Target: Number of shots on-target
5. Off-Target: Number of shots that went off-target
6. Blocked: Number of opponent team's attempts blocked by the team
7. Corners: Number of corner shots used
8. Off-sides: Number of off-side events
9. Free Kicks: Number of free-kicks used
10. Fouls Committed (FoulsC): Number of fouls committed by the team members



(a) $Topk_{ws}$ (b) TA (c) NRA

Figure 6: p-value for significance graph using Spearman coefficient



(a) $topk_{ws}$ (b) TA (c) NRA

Figure 7: p-value for significance graph using Kendall coefficient

5 Discussion

In the beginning, we analyze the performance instability when the end-user executed a request of Top_k with k=5 in the GRSS. We use dataset1 for the RS1 and the dataset2 for the RS2. We choose parameters to be maximized and others to be minimized. We evaluate with a scenario of identifying k = 5 Top_k objects over user-specified preferences on four dimensions on a range of ten dimensions. After that, we investigate if there exists a significant correlation between the parameters. The test was applied while considering our system involves user choice for answering a bi-objective problem. Then, we evaluate the runtime of recommendation according to k variation, as can be seen in figure 3, by comparing the $Top_{k_{WS}}$ algorithm with NRA and TA algorithms, respectively, responding to the final user's requirement. According to dataset1, the results show that $Top_{k_{WS}}$ outperforms the NRA and TA algorithms. Nonetheless, while the TA has a considerable runtime expressed, the runtime of $Top_{k_{WS}}$ remains very close to the NRA than that of TA. On another side, for the dataset2 and the RS2, the k variation influence more slowly the gap between the algorithm's runtime.

The $Top_{k_{WS}}$'s runtime is still inferior to the NRA and TA algorithm. We noticed that the $Top_{k_{WS}}$ and NRA reach their best runtime in k=4. For studying the Spearman and Kendall metrics, we employ ten dimensions in the used datasets. The test considers all pairs of possible variables in an example request of Top_k with k=5 to both datasets of the system. For the dataset1: the

Date Match Date	Team Playing Team	Opponent Opponent Team	Goal Scored Number of goals scored by this team	Ball Possession % Amount of time ball was in control by the team	Attempts Number of attempts to score goal
17-06-2018 6% 25-06-2018 6% Other (23) 88%	France 5% Croatia 5% Other (30) 89%	France 5% Croatia 5% Other (30) 89%	0 6	25 75	3 26
14-06-2018	Russia	Saudi Arabia	5	40	13
14-06-2018	Saudi Arabia	Russia	0	60	6
15-06-2018	Egypt	Uruguay	0	43	8
15-06-2018	Uruguay	Egypt	1	57	14
15-06-2018	Morocco	Iran	0	64	13
15-06-2018	Iran	Morocco	1	36	8
15-06-2018	Portugal	Spain	3	39	8
15-06-2018	Spain	Portugal	3	61	12
16-06-2018	France	Australia	2	51	12
16-06-2018	Australia	France	1	49	4

Figure 8: an extract from dataset 2 (FIFA, 2018 (accessed August 03, 2018))

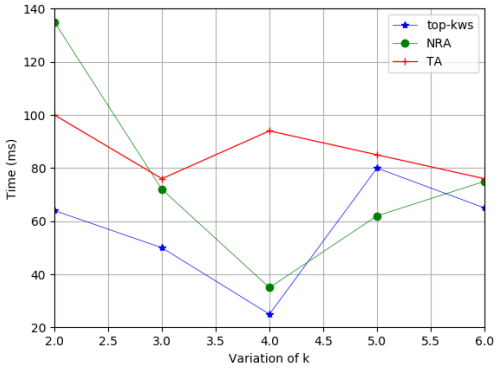
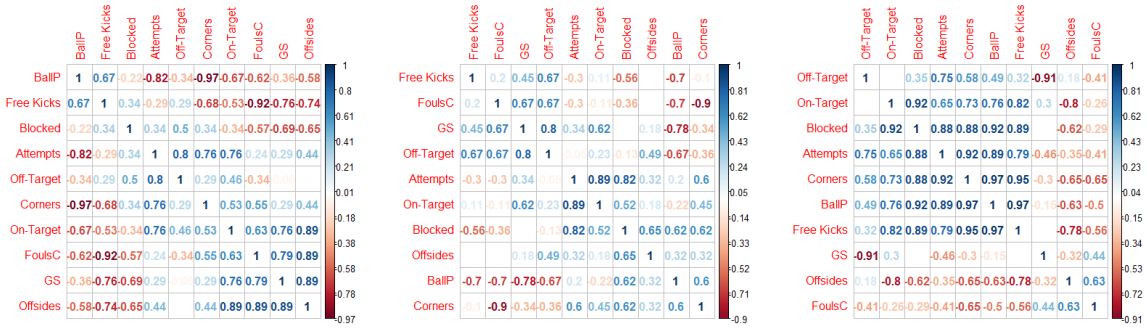
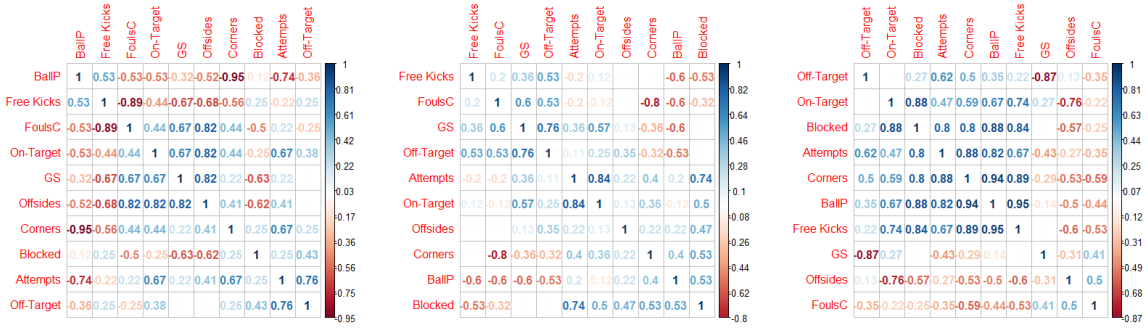


Figure 9: Response time variation of the $Topk_{ws}$, TA and NRA algorithm according to k variation using the Real dataset



(a) $Topk_{ws}$ (b) TA (c) NRA

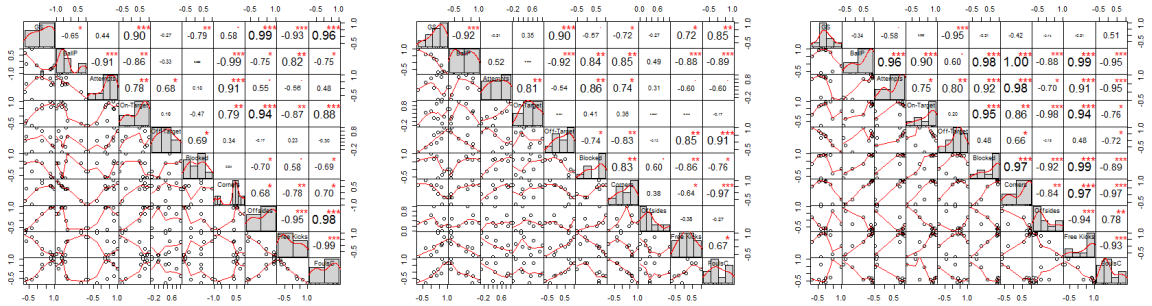
Figure 10: Spearman correlation coefficient



(a) $Topk_{ws}$ (b) TA (c) NRA

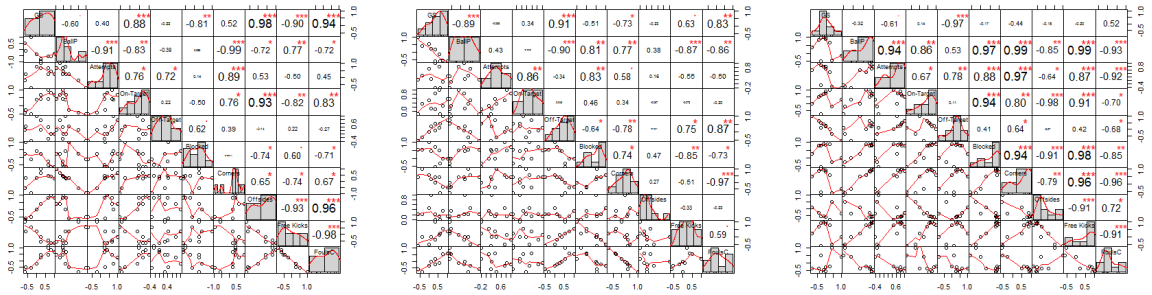
Figure 11: Kendall correlation coefficient

criteria in our formalism include seven that must be minimized: Availability, Data loss, Latency, Ongoing cost, Risk, RAM, and Response time. Besides, the criteria that must be maximized are Bandwidth, Hard Drive, and Portability. For the dataset2, we used six criteria that must be maximized: the Goal Scored, Ball possessing, Attempts, On-Target, and Blocked. Moreover, four criteria must be minimized: Off-Target, Corners, Off-sides, Free Kicks, and Fouls Committed. The Spearman coefficient estimates how strong a monotonic function could represent the relationship between two Cloud Service's parameters using dataset1 and the FIFA prediction parameters using dataset2. From figure 4, 5, 10, and 11 it can be seen that: The coefficient values are from -1 to +1. If the value is +1, the variable's relationship is perfectly monotonous and is related to an accumulating correlation. However, if the value is -1, this explains that the relationship of variables is perfectly monotonous, related to a decreasing relationship. However, when the value is equal to 0, it indicates that the variables are not related. On the other hand, the legend to the correlogram's right shows the correlation coefficients and their intensity by colors. According to the presented graphs, the comparison between the three algorithms using the Spearman metric shows a more monotonous relationship according to $Top_{k_{WS}}$, then TA and NRA. For example, the comparison based on the correlation study of the Bandwidth and latency presented in Figure 10 a, b and c, Showed respectively Spearman Correlation Coefficient (CCS) of the $Top_{k_{WS}}$ is -0.9, which is higher than TA (CCS) = -0.3, and NRA (CCS) = -0.7. In which the CCS of $Top_{k_{WS}}$ is, in general, more significant than that of TA and NRA,



(a) $Topk_{ws}$ (b) TA (c) NRA

Figure 12: p-value for significance graph using Spearman coefficient



(a) $Topk_{ws}$ (b) TA (c) NRA

Figure 13: p-value for significance graph using Kendall coefficient

which explained by a more monotonicity relationship between Cloud Computing QoS criteria while using the Top_{kWS} algorithm than the other algorithms. The same remark is given when we deal with the Kendall Correlation Coefficient showed in Figure 11. The same remark is for the Kendall metric. We noticed that Kendall metrics deal with the concordance and the discordance relationship between the parameters. While Spearman's metric, as mentioned above, study the monotonicity relationship. For more clarity of the use of correlations coefficients matrix, we present an example of the Correlation matrix of Spearman and Kendall coefficients presented respectively in Table 2 and 3 between the 2018 FIFA parameters for a query of the top-5 man of the match using Top_{kWS}. Furthermore, to completely analyze the given results, the Top_{kWS}'s superiority was improved by the p-value (Wasserstein, Lazar et al., 2016) of Spearman and Kendall metrics, presented in figures 6, 7, 12, and 13 in which The histograms of the variables shown on the diagonal. And the asterisks indicate the significance levels of the correlations. Moreover, the aforementioned statistical parameter shows the level of significance of the correlation coefficients. It is marked on the graph by "$*$", to indicate $P < 0.05$, two stars to indicate $P < 0.01$, and three stars were practiced to indicate $P < 0.001$. Whether for the coefficients: ρ of Spearman, or the τ of Kendall.

Meanwhile, the test is applied to our approach's correlation results compared to the given results by TA and NRA algorithms according to dataset 1 and 2

Table 2: Example of Spearman's correlation coefficients between the criteria of the 2018 FIFA dataset based on $Topk_{ws}$ using k=5.

	GS	BallP	Attempts	On-Target	Off-Target	Blocked	Corners	Offsides	Free Kicks	FoulsC
GS	1.00	-0.36	0.29	0.76	-0.06	-0.69	0.29	0.89	-0.76	0.79
BallP	-0.36	1.00	-0.82	-0.67	-0.34	-0.22	-0.97	-0.58	0.67	-0.62
Attempts	0.29	-0.82	1.00	0.76	0.80	0.34	0.76	0.44	-0.29	0.24
On-Target	0.76	-0.67	0.76	1.00	0.46	-0.34	0.53	0.89	-0.53	0.63
Off-Target	-0.06	-0.34	0.80	0.46	1.00	0.50	0.29	0.00	0.29	-0.34
Blocked	-0.69	-0.22	0.34	-0.34	0.50	1.00	0.34	-0.65	0.34	-0.57
Corners	0.29	-0.97	0.76	0.53	0.29	0.34	1.00	0.44	-0.68	0.55
Offsides	0.89	-0.58	0.44	0.89	0.00	-0.65	0.44	1.00	-0.74	0.89
Free Kicks	-0.76	0.67	-0.29	-0.53	0.29	0.34	-0.68	-0.74	1.00	-0.92
FoulsC	0.79	-0.62	0.24	0.63	-0.34	-0.57	0.55	0.89	-0.92	1.00

Table 3: Example of Kendall's correlation coefficients between the criteria of the 2018 FIFA dataset based on $Topk_{ws}$ using k=5.

	GS	BallP	Attempts	On-Target	Off-Target	Blocked	Corners	Offsides	Free Kicks	FoulsC
GS	1.00	-0.32	0.22	0.67	0.00	-0.63	0.22	0.82	-0.67	0.67
BallP	-0.32	1.00	-0.74	-0.53	-0.36	-0.12	-0.95	-0.52	0.53	-0.53
Attempts	0.22	-0.74	1.00	0.67	0.76	0.25	0.67	0.41	-0.22	0.22
On-Target	0.67	-0.53	0.67	1.00	0.38	-0.25	0.44	0.82	-0.44	0.44
Off-Target	0.00	-0.36	0.76	0.38	1.00	0.43	0.25	0.00	0.25	-0.25
Blocked	-0.63	-0.12	0.25	-0.25	0.43	1.00	0.25	-0.62	0.25	-0.50
Corners	0.22	-0.95	0.67	0.44	0.25	0.25	1.00	0.41	-0.56	0.44
Offsides	0.82	-0.52	0.41	0.82	0.00	-0.62	0.41	1.00	-0.68	0.82
Free Kicks	-0.67	0.53	-0.22	-0.44	0.25	0.25	-0.56	-0.68	1.00	-0.89
FoulsC	0.67	-0.53	0.22	0.44	-0.25	-0.50	0.44	0.82	-0.89	1.00

6 CONCLUSIONS

, Whether in sports management or Cloud Computing Service Selection, the improved approach based on the RS1 and RS2 shows the significant performance of the given runtime by $Top_{k_{WS}}$ algorithm compared to TA and NRA algorithm to respond to user requirements. Moreover, We try to obtain a trade-off between a good runtime and a good quality of found results. It is also worth mentioning the overall improvement of the GRSS as a general solution. Consequently, the quality of our approach results is considered engaging according to runtime evaluation and four correlation metrics. In our future work, We aim to parallelized this algorithm by utilizing it in a distributed context based on Hadoop and Spark while using more Big datasets and combining it with deep learning processing in recommendation and prediction steps.

References

Abourezq, M. and Idrissi, A. 2014a. A cloud services research and selection system, *Multimedia Computing and Systems (ICMCS), 2014 International Conference on*, IEEE, pp. 1195–1199.

Abourezq, M. and Idrissi, A. 2014b. Introduction of an outranking method in the cloud computing research and selection system based on the skyline, *Research Challenges in Information Science (RCIS), 2014 IEEE Eighth International Conference on*, IEEE, pp. 1–12.

Adomavicius, G. and Tuzhilin, A. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, *IEEE transactions on knowledge and data engineering* **17**(6): 734–749.

Amagata, D., Hara, T. and Onizuka, M. 2018. Space filling approach for distributed processing of top-k dominating queries, *IEEE Transactions on Knowledge and Data Engineering* **30**(6): 1150–1163.

Burer, S. 2012. Robust rankings for college football, *Journal of Quantitative Analysis in Sports* **8**(2).

El handri, K. E. and Idrissi, A. 2020a. Comparative study of topk based on fagin's algorithm using correlation metrics in cloud computing qos, *International Journal of Internet Technology and Secured Transactions* **10**(1-2): 143–170.

El handri, K. and Idrissi, A. 2019. Étude comparative de topk basée sur l'algorithme de fagin en utilisant des métriques de corrélation dans la qualité de service de cloud computing., *EGC*, pp. 359–360.

El Handri, K. and Idrissi, A. 2020b. Parallelization of top_k algorithm through a new hybrid recommendation system for big data in spark cloud computing framework, doi: 10.1109/jsyst.2020.3019368, *IEEE Systems Journal* .

Fagin, R., Kumar, R. and Sivakumar, D. 2003. Comparing top k lists, *SIAM Journal on discrete mathematics* **17**(1): 134–160.

Fagin, R., Lotem, A. and Naor, M. 2003. Optimal aggregation algorithms for middleware, *Journal of computer and system sciences* **66**(4): 614–656.

FIFA, S. 2018 (accessed August 03, 2018). *FIFA predect man of the match 2018*.
URL: *Statistics FIFA-2018, https:FIFA predect-2018, //www.kaggle.com/mathan/fifa-2018-match-statistics*

Gil, R. A., Johanyák, Z. C. and Kovács, T. 2018. Surrogate model based optimization of traffic lights cycles and green period ratios using microscopic simulation and fuzzy rule interpolation, *Int. J. Artif. Intell* **16**(1): 20–40.

Hwang, S.-w. and Chang, K. C.-c. 2007. Optimizing top-k queries for middleware access: A unified cost-based approach, *ACM Transactions on Database Systems (TODS)* **32**(1): 5–es.

Idrissi, A. and Abourezq, M. 2014. Skyline in cloud computing., *Journal of Theoretical & Applied Information Technology* **60**(3): 1992–8645.

Idrissi, A., El handri, K., Rehioui, H. and Abourezq, M. 2016. Top-k and skyline for cloud services research and selection system, *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies*, ACM, p. 40.

Kalavagattu, A. K., Das, A. S., Kothapalli, K. and Srinathan, K. 2011. On finding skyline points for range queries in plane., *CCCG*.

Lai, Y.-J., Liu, T.-Y. and Hwang, C.-L. 1994. Topsis for modm, *European journal of operational research* **76**(3): 486–500.

Liu, J., Xiong, L., Pei, J., Luo, J. and Zhang, H. 2015. Finding pareto optimal groups: Group-based skyline, *Proceedings of the VLDB Endowment* **8**(13): 2086–2097.

Liu, J., Xiong, L., Pei, J., Luo, J., Zhang, H. and Yu, W. 2019. Group-based skyline for pareto optimal groups, *IEEE Transactions on Knowledge and Data Engineering* .

Marung, U., Theera-Umpon, N. and Auephanwiriyakul, S. 2016. Top-n recommender systems using genetic algorithm-based visual-clustering methods, *Symmetry* **8**(7): 54.

Osaba, E., Del Ser, J., Sadollah, A., Bilbao, M. N. and Camacho, D. 2018. A discrete water cycle algorithm for solving the symmetric and asymmetric traveling salesman problem, *Applied Soft Computing* **71**: 277–290.

Ott, E., Hayashi, N. and Fukuda, M. 2007. Method and system for using smart tags and a recommendation engine using smart tags. US Patent App. 11/424,966.

Precup, R.-E. and David, R.-C. 2019. *Nature-inspired optimization algorithms for fuzzy controlled servo systems*, Butterworth-Heinemann.

Purcaru, C., Precup, R.-E., Iercan, D., Fedorovici, L.-O., David, R.-C. and Dragan, F. 2013. Optimal robot path planning using gravitational search algorithm, *International Journal of Artificial Intelligence* **10**(S13): 1–20.

Saeed, M., Saqlain, M. and Riaz, M. 2019. Application of generalized fuzzy topsis in decision making for neutrosophic soft set to predict the champion of fifa 2018: A mathematical analysis, *Journal of Mathematics (ISSN 1016-2526)* **51**(8): 111–126.

Saporta, G. 2006. *Probabilités, analyse des données et statistique*, Editions Technip in Paris, French.

Tiakas, E., Valkanas, G., Papadopoulos, A. N., Manolopoulos, Y. and Gunopulos, D. 2016. Processing top-k dominating queries in metric spaces, *ACM Transactions on Database Systems (TODS)* **40**(4): 23.

Tzeng, G.-H. and Huang, J.-J. 2011. *Multiple attribute decision making: methods and applications*, CRC press.

Vaziri, B., Dabadghao, S., Yih, Y. and Morin, T. L. 2018. Properties of sports ranking methods, *Journal of the operational research society* **69**(5): 776–787.

Wasserstein, R. L., Lazar, N. A. et al. 2016. The asa's statement on p-values: context, process, and purpose, *The American Statistician* **70**(2): 129–133.

Wheatcroft, E. 2020. Forecasting football matches by predicting match statistics, *arXiv preprint arXiv:2001.09097* .

Wordcup, N. 2018 (accessed April 15, 2020). *worldcup news*.
URL: *worldcup news russia-2018, https://www.fifa.com/worldcup/news/russia-2018-most-engaging-fifa-world-cup-ever*