ACTA INFOLOGICA 2023;7(1):209-228
dergipark.org.tr/acin

İSTANBUL UNIVERSITY PRESS

DOI: 10.26650/acin.1198925

RESEARCH ARTICLE

A Modular Efficiency Determination Formula for Information Retrieval Evaluations and Optimizations

Bilgi Erişim Değerlendirmeleri ve Optimizasyonları İçin Modüler Bir Verimlilik Belirleme Formülü

Veli Özcan Budak[1]

ABSTRACT

The notion of efficiency has typically been associated with the efficiency of systems rather than users in Information Retrieval (IR) literature. In the usability literature, on the other hand, this notion is defined from a user-based perspective, corresponding to how long a user accomplishes a task. Despite this, the common aim for both has to do with the efficient use of time. This study examines the efficiency notion in the IR literature from a user-based efficiency window in the usability literature. In the present study, a modular efficiency determination formula (MEDEF) to create different efficiency indicators by focusing on IR system evaluations and optimizations from the usability perspective is proposed. The MEDEF can be thought of as an efficiency indicator creator based on both effectiveness metrics and efficiency indicators already used in IR studies. In the scope of this study, eight MEDEF-based efficiency indicators were created and compared to several baseline efficiency indicators already used in IR studies. While the study's first aim is to reveal how consistent the MEDEF-based indicators are and whether these indicators are more successful/reliable than the baselines, the second is to set an example of the usage of efficiency indicators in evaluations of IR systems from a usability perspective. General findings from interactive user behaviour for one month show that the MEDEF-based indicators outperform the baseline indicators and further strengthen the reflections in the baseline indicators. Several usage scenarios regarding the potential of the MEDEF are also shared and discussed in the scope of the study.

Keywords: Human-Computer Interaction, Efficiency Indicator, Session Abandonment, Interactive Information Retrieval

ÖZ

Verimlilik kavramı, Bilgi Erişim (BE) literatüründe temel olarak kullanıcılardan ziyade sistemlerin verimliliği ile ilişkilendirilmiştir. Kullanılabilirlik literatüründe ise bu kavram, bir kullanıcının bir görevi ne kadar sürede tamamladığına karşılık gelen kullanıcı tabanlı bir bakış açısıyla tanımlanır. Yine de, ortak amaç her iki literatür için de zamanı verimli kullanmaktır. Bu çalışma, BE literatüründeki etkinlik kavramını, kullanılabilirlik literatüründeki kullanıcı tabanlı etkinlik penceresinden incelemektedir. Bu çalışmada, kullanılabilirlik perspektifinden BE sistem değerlendirmelerine ve optimizasyonlarına odaklanarak farklı verimlilik göstergeleri oluşturmak için modüler bir verimlilik belirleme formülü (MEDEF) önerilmiştir. MEDEF, BE çalışmalarında hâlihazırda kullanılan etkililik metriklerine ve verimlilik göstergelerine dayalı bir verimlilik göstergesi üreticisi şeklinde düşünülebilir. Bu çalışma kapsamında, sekiz MEDEF tabanlı verimlilik göstergesi oluşturulmuş ve hâlihazırda BE çalışmalarında kullanılan birkaç temel verimlilik göstergesiyle karşılaştırılmıştır. Çalışmanın ilk amacı, MEDEF temelli göstergelerin ne kadar tutarlı olduğunu ve bu göstergelerin mevcut temel göstergelere göre daha başarılı/güvenilir olup olmadığını ortaya koymak iken, ikincisi, kullanılabilirlik açısından BE sistemlerinin değerlendirmelerinde verimlilik göstergelerinin kullanımına bir örnek oluşturmaktır. Bir aylık etkileşimli kullanıcı davranışlarından elde edilen genel bulgular, MEDEF tabanlı göstergelerin temel göstergelerden daha iyi performans gösterdiğini ve temel göstergelerdeki yansımaları daha da güçlendirdiğini göstermiştir. MEDEF'in potansiyeline ilişkin çeşitli kullanım senaryoları da çalışma kapsamında paylaşılmakta ve tartışılmaktadır.

Anahtar Kelimeler: İnsan-Bilgisayar Etkileşimi, Verimlilik Göstergesi, Oturum Terk Etme, Etkileşimli Bilgi Erişim

[1](Assist. Prof.), Bandırma Onyedi Eylül University, Gonen Vocational School, Balikesir, Turkiye

ORCID: V.Ö.B. 0000-0002-0960-0542

Corresponding author:
Veli Özcan BUDAK
Bandırma Onyedi Eylül University, Gonen Vocational School, Balikesir, Turkiye
E-mail address: veliozcanbudak@gmail.com

Submitted: 03.11.2022
Revision Requested: 25.05.2023
Last Revision Received: 26.05.2023
Accepted: 26.05.2023
Published Online: 19.06.2023

Citation: Budak, V.Ö. (2023). A modular efficiency determination formula for information retrieval evaluations and optimizations. *Acta Infologica, 7*(1), 209-228. https://doi.org/10.26650/acin.1198925

1. INTRODUCTION

Evaluations in Information Retrieval (IR) studies are made mostly by following the Cranfield paradigm and focusing on the effectiveness of systems/models. Two elements underlie the basis of these kinds of evaluations. While reference collections with predetermined information needs correspond to the first element, several evaluation metrics constitute the other. This evaluation method has become indispensable for IR studies and even for those defined as interactive. On the other hand, the efficiency notion can be shown as the other type of evaluation. This notion is mainly associated with the efficiency of IR systems rather than that of users. Two performance elements, throughput and latency, constitute the center of efficiency investigations (Croft, Metzler, and Strohman, 2009; Büttcher, Clarke, and Cormack, 2010), and the main focus is on how quickly a system responds and the sufficiency of system resources (Zhai and Massung, 2016). In other words, efficiency is predominantly associated with the subject of "time." Regardless of the evaluation type, both evaluations intrinsically consist of system-based investigations rather than user-based ones.

When focusing on another research area, usability, it can be seen that the notion of "efficiency" also exists in this area but with a different definition. This notion is defined from a user-based perspective, corresponding to how long it takes for a user to accomplish a task (Nielsen, 1993; Frøkjær, Hertzum, and Hornbæk, 2000; Hornbæk, 2006; Rubin and Chisnell, 2008; Rosenzweig, 2015). Nevertheless, the main subject does not change when compared to the IR perspective: "time." In other words, both research areas focus on the efficient use of time. This study examines the concept of efficiency in the IR literature from a user-based efficiency window in the usability literature, meaning that the focus is on user efficiency rather than on system efficiency.

Logs of search interactions have been intensively utilized for different purposes in IR studies because they consist of natural user behaviors. While Jiang, Leung, Yang, and Ng (2015) and Vidinli and Ozcan (2016) utilized logs for query suggestion purposes, Joachims (2002) and Agichtein, Brill, and Dumais (2006) used them for search result improvement. Liu, Liu, and Belkin (2014) and Kim, Hassan, White, and Zitouni (2014), on the other hand, used logs for the exploration of user behavior patterns. Users who try to satisfy their information needs leave several types of traces (implicit data) in the background while searching. These traces also hold some valuable indicators regarding the time notion. "Dwell time," "time to first click," and "time to last click" can be given as examples of these types of indicators. These indicators have been considered with different definitions/namings in IR studies, such as "user engagement" by Singla and White (2010), "search satisfaction" by Hassan (2012), Kim, Hassan, White, and Zitouni (2014), and Liu et al. (2015), "choice overload" by Beierle, Aizawa, Collins, and Beel (2020), "implicit measures of user interest" by Fox et al. (2005), and "interest detectors" by Claypool, Brown, Le, and Waseda (2001). When considering them from the usability perspective, these indicators carry signs regarding user efficiency even though the namings are different. On the other hand, even if we put this differentiation aside, it is evident that these indicators have not been utilized individually in evaluating IR systems, as explained in Section 2. At this point, two questions appear that need to be answered: (1) How reliable or useful are these indicators when they are not used for the purpose of evaluation? (2) Is it possible to evaluate IR systems only through these indicators? While the first question regards the usage of these indicators already in IR studies, the second regards whether these indicators can be used to evaluate IR systems from the usability perspective. This study seeks answers to these questions and, in addition to the mentioned efficiency indicators, proposes a modular efficiency determination formula (MEDEF) to create different types of user-based efficiency indicators.

In the scope of the present study, users' search interactions were first recorded through the search modules of different department websites of a university in Turkey. Secondly, these websites (23 in total) were separated into six groups based on users' session abandonment behaviors, and fifteen propositions were created regarding the IR performances of these groups. Lastly, eight efficiency indicators created by following the MEDEF and four baseline indicators already used in IR studies were compared statistically to investigate how they successfully reflect group performances based on the created propositions. Apart from the statistical comparisons, several examinations were also made with Machine Learning (ML)-based approaches. The study's first aim is to reveal how consistent the MEDEF-based indicators are and whether these indicators are more successful/reliable than the baselines. The second is to set an example of the usage of efficiency indicators in evaluations of IR systems from a usability perspective.

The rest of the paper is organized as follows. Studies made with efficiency indicators are shared in Section 2. Section 3 explains the MEDEF's features and the method followed in the present study. The study's findings are shared in Section 4. While the results are discussed in Section 5, the contributions of the present study and several usage scenarios of the MEDEF for future work are shared in Section 6.

2. RELATED WORK

In this section studies which have benefited from time-based efficiency indicators are shared.

Hassan, Song, and He (2011) presented a prediction model based on the ML concept to determine user satisfaction in the information-seeking process. While creating the model, the authors benefited from the dwell time indicator, and the model's success in estimating user satisfaction and improving search results was revealed. Lee, Teevan, and de la Chica (2014) conducted a study to characterize user search behavior by utilizing the time to first click and dwell time indicators. The authors made several suggestions for search result improvement in the study results. Based on the ML concept, Arguello (2014) created a model by using the dwell time indicator and several implicit data types to predict search task difficulty. In the study's result, several suggestions regarding utilizing different types of data/indicators in creating a predictive model were made by the author. Similarly, Liu, Liu, and Belkin (2014) also investigated users' behavioral differences regarding search task difficulties in the information-seeking process. The authors created a prediction model utilizing several types of dwell time indicators to determine search task difficulties. In their study, Kim, Hassan, White, and Zitouni (2014) created a model based on the ML concept using the dwell time indicator and tried to predict search satisfaction. In another study using the dwell time indicator an attempt was made to categorize search sessions using the ML concept to show whether they reflected struggling or exploring (Hassan, White, Dumais, and Wang, 2014). Balakrishnan and Zhang (2014) proposed a model benefiting from the dwell time indicator to improve search results. The authors stated that the proposed model revealed significant improvements. A model using the time-to-first click, dwell time, and several different indicators was created by Arkhipova, Grauer, Kuralenok, and Serdyukov (2015) to predict unsuccessful search sessions. The authors investigated users' search engine switching behaviors in the study in which an evaluation metric was also presented for A/B testing studies.

In a study by Borisov, Markov, de Rijke, and Serdyukov (2016), the authors tried to explore behavioral patterns on times between user actions using the time to first and last click and dwell time indicators and proposed a prediction model to be used for personalized IR improvements. The study results indicated that successful findings were reached. Alhabashneh, Iqbal, Doctor, and James (2017) proposed a fuzzy-based mechanism using the dwell time indicator and different implicit data types to estimate document relevance. The authors indicated that the proposed mechanism showed successful prediction performance regarding document relevance. In another study focused on system-based efficiency, a framework proposal was made by Makkar and Kumar (2018) to improve search results by utilizing the dwell time indicator. The authors achieved successful results with the proposed framework. Another study focusing on difficulties in making decisions while searching research articles was made by Beierle, Aizawa, Collins, and Beel (2020). In the study, in which one of the indicators was the time to first click indicator, the authors revealed findings regarding users' behavior patterns. Sarkar, Mitsui, Liu, and Shah (2020) conducted a study focusing on determining users who needed help finding information. The authors created a successful prediction model with the logistic regression classifier, utilizing the dwell time indicator.

When we consider the studies above, it can be seen that two main types of investigation exist: revealing users' behavioral patterns or using efficiency indicators for optimization purposes. As evaluations were made following the Cranfield paradigm regardless of the study types, these studies neither focus on how reliable the efficiency indicators are among each other nor consist of evaluations from users' perspectives. The present study sets an example for evaluating IR systems from a user-based approach. Moreover, it sheds light on the trustworthiness of efficiency indicators and how these indicators can be utilized more successfully through the MEDEF for both IR system evaluations and optimizations.

3. METHOD

This section first explains the components of the data collection process. Secondly, the efficiency indicators that have already been used in IR studies, and the efficiency indicators created by following the proposed formula in the present study, are

explained by being combined with the data preprocessing step. After this, the research questions and how the evaluations were carried out are shared. Lastly, the section concludes by explaining the present study's limitations.

3.1. Data Collection Process

User search-visit interactions, gathered over the period of one month from the textual search modules of different department websites of a university in Turkey , underlie the center of this study. The search modules serve with the boolean model (the same infrastructure) and work integrated with the websites. As each website addresses different information needs, the contents announced differ, meaning each search module can be considered a tool that helps users satisfy different information needs. The interaction data were collected from 66 departments. A relational database with two tables (*queries* and *visits*) was used to record the interactions with these modules during the data collection. The structure of the tables is given in Table 1.

Table 1
The used tables in recording search interactions

	The *queries* table	**The *visits* table**
Attributes	id	id
	department_id	query_id
	query	page_id
	query_length	rank
	dwell_on_SERP*	visit_time
	query_time	
	anonymized_ip	

* The dwell time on the Search Engine Result Page (SERP). This value was recorded in the form of seconds.

As seen in Table 1, both tables were created with a simple design to collect basis interaction data. A record was added to the *queries* table whenever a user searched with a query for information in the data collection process. After users were shown relevant pages and their snippets found by modules, the users had two options: clicking a suggestion on the SERP or abandoning the session/query. While a record regarding the visit behavior was added to the *visits* table for users who preferred the first option, no record was added to the table for the other option, and session states were regarded as being abandoned.

Two different values, query length (QL) and dwell time on the SERP (DwSERP), were also recorded in addition to query time (QT) in each search session (Table 1; the *queries* table) to be utilized in the created indicators. While QL is how many unique words a query term has, DwSERP corresponds to how much time a user spends on the SERP in a search session. DwSERP was recorded at 0.1-second intervals after starting a search session by making a query (a Javascript code was utilized for this process). For the *visits* table, the main values utilized in the indicators are rank and visit time (VT). While the rank value corresponds to what order a clicked web page (document) suggestion is on the SERP, VT indicates the first contact timestamp between users and the clicked documents.

Throughout the data collection process, while the total number of search interactions for all websites was detected as 29,545, the total number of unique queries was observed as 3,402. How these collected data were preprocessed is explained in the next section.

3.2. Data Preprocessing and Efficiency Indicators

In the preprocessing phase, each query was first considered as a session of 30 minutes. Let us assume that a user made a query and started a session. If that user sent another query with the same terms before the opened session ended, that user's visit interactions from the second query were treated as if the user had not sent the second query, meaning that only visit interactions of the user continued to be recorded under the opened session instead of a new session being opened. A new session would be started if that user had sent the second query with different terms, even though the intent was identical. On the other hand, during preprocessing, which session belongs to which user was determined according to users' anonymized_IP

values (Table 1), and a session in which no visit interaction was observed was considered abandoned. While tracking users' visit interactions on the same session, if users clicked the same suggestion for the second/third/fourth time, only the first visit was taken into account (this action was excluded for only one indicator, as explained in Section 3.2.2). In the direction of these procedures, the preprocessing of collected data was carried out in three steps incrementally. These steps are explained below:

Step 1: Users who had made a query in the past six months were excluded from the collected data based on their anonymized IP addresses. The reason for this is to avoid fallacious data that might arise from users and that could possibly have prior information gathered from a past search interaction with the search modules regarding their continuing information needs. In other words, the situation of a user searching with a query with search modules in the past indicates that that user has an experience from the past regarding responses from the query sent in the past. In this case, if the user prefers to use the same query terms for the second time and departments shared no content regarding the user intent until this second query, the used search module will probably suggest the same documents as in the first query. In this direction, the user's decision time will quite likely be very short because of past experience. As this case can misguide the analysis, users from the past six months were excluded, meaning that users who belonged to the month of the data collection process were considered users who were interacting with the system for the first time.

The second action in this step was to exclude queries with no result from the collected data.

Step 2: Except for the search interactions whose DwSERP value was between one and 60 seconds, all interactions were treated as outliers. This action was taken in order to focus only on search interactions made with a high concentration. For the same purpose, the second action was to eliminate the interactions where the first click time (time to first click - TTFC) was higher than 30 seconds in sessions that were not abandoned. The last process in this step was to exclude websites with search sessions of less than 100.

Step 3: The number of cases of session abandonment pertaining to each website was examined after the second step. Websites in which all sessions had been abandoned were excluded from the dataset. Lastly, interactions made by the same users in different departments were removed based on users' anonymized IP addresses in order that statistical analyses be properly carried out. In other words, if a user had searched on both A department and B, only one of these interactions was kept, the other being removed. Thus, interactions in all departments were isolated from each other.

The collected data were preprocessed based on the above steps. The search interaction data, which totaled 29,545 from 66 department websites before preprocessing, were reduced to 11,228 from 23 department websites with 1691 unique query terms. Three different indicator classes were created. While the first two classes correspond to the indicators that reflect natural user behavior and have already been used in IR studies, the third consists of the indicators, which again reflect natural user behavior, created for this study. In contrast to studies that mostly focused on IR system performances/optimizations, in this study, all indicators are considered from the usability perspective, and natural user search behaviors constitute the center of investigations. From this point, the first indicator class is named "Guidance Indicator"; the second class is defined as "Baseline Efficiency Indicators"; the third class is named "MEDEF-Based Efficiency Indicators."

3.2.1. Guidance Indicator

Users' Session Abandonment (SAb) behaviors are a sign of unsuccessful sessions (Liu, Gwizdka, and Liu, 2010), and this type of user behavior can be utilized to evaluate IR systems (Diriye, White, Buscher, and Dumais, 2012). In this study, because the SAb behavior allows detection of how well IR systems serve users in satisfying their information needs from a generic perspective, it is selected as the guidance indicator. To this end, firstly, the IR performance of each website was characterized by the users' SAb behavior. Secondly, which website outperformed the other was determined by examining the percentage of sessions with no abandonment on each website. Afterward, to evaluate the reflectiveness performance of efficiency indicators, the websites that performed similarly to each other in satisfying user information needs were grouped based on the percentage of sessions with no abandonment. In this direction, six groups were created from 23 departments (Table 2).

Table 2
The groups of departments based on their performances

Name of the groups	The number of departments in the groups	Total interaction	Rule* (>X and <Y)	% of sessions with no abandonment
A	7	1296	>0 and <10	3,8
B	4	1317	>=10 and <20	15,1
C	5	7210	>=20 and <30	26,3
D	3	451	>=30 and <40	32,6
E	3	692	>=40 and <50	42,9
F	1	262	>=50	53,1

***** The percentage of sessions with no abandonment is higher than X and less than Y.

Table 2 indicates that group F is more successful than group E, that group E is more successful than group D, and so on. Fifteen propositions (6x5/2; 6 is the total number of the groups) were created to investigate the indicators' reflectiveness performance based on Table 2 (the statistical proof of the propositions is explained in Section 4). The statistical examinations were carried out by comparing the efficiency indicators between each other and focusing on how many propositions an indicator can reflect. Different ML-based examinations were also carried out using the created groups and the efficiency indicators together.

3.2.2. Baseline Efficiency Indicators

In the present study, four different efficiency indicators that have already been used in IR studies were chosen as baselines and explained below:

DwSERP

As described before, the length of time a user spends on the SERP in a search session corresponds to DwSERP. This indicator is inversely proportional to user satisfaction in the information-seeking process. High DwSERP values, which could arise from useless information sources in results or ambiguous query terms that users express, could point out negative experiences (Aula, Khan, and Guan, 2010; Sarkar, Mitsui, Liu, and Shah, 2020). Similarly, it has been stated that having difficulties in satisfying information needs causes more cognitive effort and longer DwSERP (Kuhar and Merčun, 2022). As DwSERP can reflect user satisfaction and, accordingly, how well IR systems serve users, it is chosen as the first baseline indicator. The last point that needs to be clarified for this indicator is that the DwSERP values were separated based on whether a session was abandoned or not in the analyses. The first indicator consists only of the DwSERP values of sessions with no abandonment.

DwSERP of abandoned sessions (DwSERP_Ab)

In addition to DwSERP, another baseline indicator was created from abandoned sessions' DwSERP values. As in DwSERP, we believe that DwSERP_Ab can also offer clues as to how close an IR system is to meeting information needs. In this direction, the tested assumption in the analyses was that the shorter DwSERP_Ab is, the more likely IR systems are close to satisfying information needs.

TTFC

TTFC is the length of time that elapses between the point at which a session starts upon a user sending a query and the point at which that user clicks on one of the pages on the SERP for the first time. In the study by Radlinski, Kurup, and Joachims (2008), it was found that TTFC was correlated with search success, meaning that the quality of results on the SERP decreases when TTFC increases. The faster users find relevant information sources the shorter TTFC is. From this perspective, it can be stated that as users do not have many difficulties when making the first decision, search interactions result in a positive experience. TTFC was chosen as the third indicator to investigate whether it can reflect user behavior in the propositions created through the SAb behaviors. While calculating this value, the QT value was subtracted from the first VT value observed for each session (Table 1).

Time to last click (TTLC)

TTLC is the time that elapses from the moment a session starts upon a user sending a query and the moment the user clicks one of the pages on the SERP for the last time in that active session. When users clicked the same suggestion on the SERP more than once, it was stated that only the first visit was considered except for one indicator. While the first click was recorded for the TTFC indicator, the others were utilized for the TTLC indicator regardless of being the same or different suggestions. The assumption for this indicator is the same as the TTFC indicator. In other words, shorter TTLCs reflect that users are able to meet their information needs in a short time, which indicates a more positive experience (Radlinski, Kurup, and Joachims, 2008).

3.2.3. MEDEF-Based Efficiency Indicators

This section is organized into two parts. While the formula used in creating efficiency indicators is explained in the first part, the second shares eight different indicators created by following the proposed formula.

The MEDEF

The proposed formula is used to give a value to each unique session that is not abandoned. It is calculated with the combination of three metrics: "Ambiguity Reward," "Punishment," and "Effectiveness," and can be thought of as an indicator that consists of both an efficiency and effectiveness metric together (Equation 1).

$$MEDEF = \underbrace{\frac{1}{QL}}_{\text{Ambiquity Reward}} \cdot \underbrace{\frac{1}{baseline\_indicator}}_{\text{Punishment}} \cdot effectiveness \qquad (1)$$

Ambiguity Reward: The Query Expansion (QE) technique has been intensively used in IR studies (Colace, De Santo, Greco, and Napoletano, 2015; Nie et al., 2016; Singh and Sharan, 2017; Nasir, Varlamis, and Ishfaq, 2019) to expand users' initial queries and thus alleviate the burden of finding relevant sources associated with user queries on these systems. Adding similar but differently expressed terms to initial terms to increase the possibility of finding relevant sources by IR systems underlies the QE technique. As QL supports IR systems (Belkin et al., 2003; Kelly and Fu, 2007), reciprocal QL is used as a reward in the MEDEF formula. In other words, when QL increases, the ambiguity decreases, and this increase helps IR systems understand search intents more easily. That is why the more words a query term has, the less reward it gets per session. This approach especially rewards sessions that start with ambiguous queries. In addition, it should be mentioned that no QE technique was used in the present study; rewards were given based on users' natural term selections.

Punishment: When we consider the baseline indicators in the previous section, it can be seen that all of them have a common logic, which is that the increase in an indicator points out dissatisfaction in terms of IR experience. To this end, the reciprocal value of efficiency indicators is utilized to penalize sessions.

Effectiveness: IR studies in the literature focus on whether IR systems/models can satisfy information needs successfully, and for this purpose, evaluations based on effectiveness metrics and reference collections mainly constitute the basis of these kinds of studies. Several metrics have been proposed/created from the past to the present. Gain-based metrics (CG, NCG, DCG, NDCG) by Järvelin and Kekäläinen (2000) and Järvelin and Kekäläinen (2002); Expected Reciprocal Rank (ERR) by Chapelle, Metlzer, Zhang, and Grinspan (2009); Rank-Biased Precision (RBP) by Moffat and Zobel (2008); Binary Preference (BPref) by Buckley and Voorhees (2004) and Mean Reciprocal Rank (MRR) by Voorhees (1999) can be shown as examples for effectiveness metrics. These and homologous metrics focus on determining the success of IR systems/models based on how many relevant/useful information sources are found using reference collections. However, two disadvantages exist in using effectiveness metrics for interactive IR studies. One is the nonexistence of reference collections in interactive environments, as users' information needs constantly change. The other is biased user behavior (Joachims, 2002; Joachims et al., 2005; Agichtein, Brill, and Dumais, 2006). Nonetheless, it was decided to integrate effectiveness metrics into the

formula because users' natural behaviour can take the place of explicit relevance judgments (Croft, Metzler, and Strohman, 2009; Zhai and Massung, 2016), and collecting these natural behavior data in large quantities is easy (Manning, Raghavan, and Schütze, 2008). Effectiveness metrics were not integrated into the formula to reward/penalize sessions. However, whether sessions are rewarded/penalized is left to user actions. Let us assume that the MRR metric is used, and the multiplication of the first two metrics in the formula is calculated as 0.3 for a session. If the user of that session clicks only on the second suggestion on the SERP, the MRR value will be calculated as 1/2 = 0.5, and the MEDEF value of the session will be 0.15. In the other scenario, if the user clicks two suggestions at the ranks of first and second, the MRR value will be calculated as (1/1 + 1/2) / 2 = 0.75, and the MEDEF value of the session will be 0.225. In short, user preference will decide whether sessions are rewarded/penalized.

In light of the above explanations, the last point that needs to be explained is that the increase in the MEDEF value indicates more positive user experiences.

The Created Indicators

Eight indicators with different variations were created using the combination of three baseline indicators and two effectiveness metrics (MRR and Average Precision - AP) to investigate how successful the MEDEF-based indicators perform in reflecting user behavior regarding the propositions created with the guidance indicator (Table 3).

Table 3
The created indicators based on the MEDEF

Description	Punishment	Effectiveness
MRR_DwSERP	$\frac{1}{DwSERP}$	$MRR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{rank_i}$
MRR_TTFC	$\frac{1}{TTFC}$	
MRR_TTLC	$\frac{1}{TTLC}$	
MRR_ALL	$\frac{1}{DwSERP \cdot TTFC \cdot TTLC}$	
AP_DwSERP	$\frac{1}{DwSERP}$	$AP(L) = \frac{1}{\lvert Rel \rvert}\sum_{i=1}^{n} P(i)$
AP_TTFC	$\frac{1}{TTFC}$	
AP_TTLC	$\frac{1}{TTLC}$	
AP_ALL	$\frac{1}{DwSERP \cdot TTFC \cdot TTLC}$	

Because of the nature of interactive environments, it is impossible to determine which information sources are relevant to a query. Nevertheless, users' clickthrough behaviors can shed light on this ambiguity. Clicks can be utilized in determining whether a document is relevant or not. While using the MRR metric in the indicators, as exemplified before, the MRR value of a session can be calculated considering all rank values of visits of that session. The same process is also followed for calculating each session's AP value. We can define AP for a session with a ranked list L (on the SERP) where $|L|=n$; $P(i)$ donates the precision value of a clicked document at rank i in L; and *Rel* is all relevant documents in the collection. Even though *Rel* is ambiguous in interactive environments, the calculations can be carried out by accepting that *Rel* corresponds to all documents visited by a user in a session. If we exemplify the AP calculation through the same two scenarios given for MRR, the first scenario gives us the same result again because the user only visits the second document in L: 1/2 = 0,5. As for the second, as the user visits documents at the first and second ranks, the AP value will be calculated as (1/1 + 2/2) / 2 = 1. Let us assume another scenario where that user clicks three suggestions at the second, third, and fifth ranks in the same session. The AP value will be calculated as (1/2 + 2/3 + 3/5) / 3 = 0,59.

The MRR and AP values were calculated by following the above method for each session. Afterward, the MEDEF values were determined by multiplying these values with the other two metrics for each session.

3.3. Research Questions and Evaluation Procedure

In the scope of the present study, four research questions were prepared. This section explains these questions and how the evaluations were carried out respectively.

RQ1. What are the differences between the SAb behaviors in the groups?

The reason for creating this question is first to characterize websites that performed similarly based on users' SAb behaviors. The second is to group these websites according to their performance, as in Table 2. The last is to create a set of propositions being followed as guidance regarding the IR performances of these groups of websites. The findings revealed from this question guided the investigations into the other questions.

RQ2. How do baseline efficiency indicators perform in reflecting users' SAb behavior?

This question attempts to answer how reliable the most preferred efficiency indicators in IR studies are in reflecting the service success of IR systems from a user-based window. In addition, each baseline indicator was ranked from the best to the worst on its trustworthiness.

RQ3. How do MEDEF-based efficiency indicators perform in reflecting users' SAb behavior?

The investigations into this question are in the same line as for the second question. Besides, the findings from this question were also used in comparing MEDEF-based efficiency indicators with the baselines to clarify which indicators are more reliable.

To sum up, while examining all indicators statistically, the focus was on how reliable each indicator was in consistently reflecting behavioral patterns in the propositions. The dataset utilized in the examinations consists of a total of 11,228 rows with 16 different attributes:

- session_id
- department_id
- group (refers to which group a department belongs to)
- the baseline indicators (four kinds in total)
- the MEDEF-based indicators (eight kinds in total)
- abandonment (if the session is abandoned, then this value is *true*; otherwise it is *false*)

Each row in the dataset corresponds to a unique session together with its attributes. As the prepared dataset did not fit the normal distribution for each indicator, nonparametric test methods were utilized during the examinations.

RQ4. Which prediction model created based on the type of indicators shows more successful performance?

All examinations for the other questions were carried out from a statistical perspective. Three more datasets were also created from the same dataset for this question to determine the success of indicators with an ML-based perspective. To ensure that all datasets consisted of sessions with no abandonment, the first dataset (baselines) has three baseline indicators (as independent variables) on each row with its group (as the dependent variable); the second (medef_MRR) also has three MEDEF-based indicators (independent variables; MRR_ALL was excluded) created using the MRR metric as the punishment on each row with its group (dependent variables); the third (medef_AP) has three MEDEF-based indicators (independent variables; AP_ALL was excluded) created using the AP metric as the punishment on each row with its group (dependent variables). Two supervised ML algorithms were utilized to create prediction models based on these prepared datasets: Random Forest Classifier (RF) and Decision Tree Classifier (DT). The holdout sampling was used with three different train-test separation rates (shared in Section 4.4). In addition, as the data on each dataset were unbalanced (Table 2), the stratification process

was made while sampling. Precision (P), F1 score, and Accuracy (ACC) metrics were used to assess the created models by the ML algorithms. The scikit-learn module by Pedregosa et al. (2011) was used for all examinations.

What is expected from the models is not to predict the groups perfectly but to reveal several signs that can be interpreted as directly proportional to findings from statistical examinations.

3.4. Limitations

The possibility of users' IP addresses changing after a while is the first limitation of this study because, while user interactions were organized in the preprocessing phase, these addresses were utilized. The second limitation can be stated as arising from the Javascript code used to record dwell times on the SERP. The possibility of users' browsers not supporting the code properly might cause faulty data recording. Technical problems that might arise from the server infrastructure that hosts the websites can be listed as the last limitation. The situation of users encountering this limitation in information-seeking processes might result in unfinished interactions.

4. FINDINGS

In this section, the question of whether users' search habits have changed from the past to the present is first examined, after which the descriptive statistics of indicators are shared. The section concludes with the relationship among the indicators being described and the research questions being answered.

4.1. Users' Search Habits

While the users conducted searches over the period of one month, they used query terms consisting of different numbers of words. A consideration of all sessions led to five groups being created based on how many words the query terms had (Figure 1).

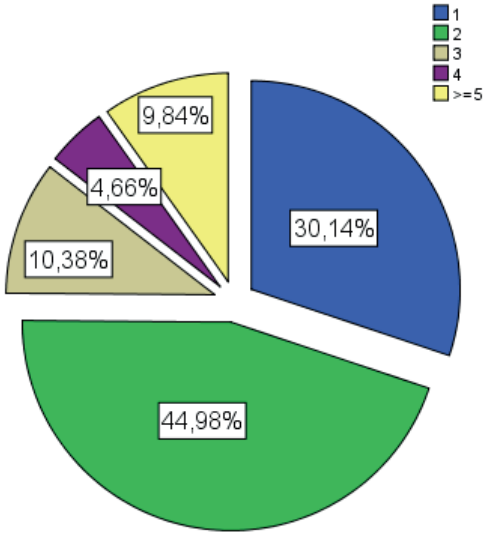

Figure 1. The groups of query terms based on the number of words

It seems that the findings from the studies by Jansen, Spink, Bateman, and Saracevic (1998), Jansen, Spink, and Saracevic (2000), and Jansen and Spink (2005) have maintained their validity in that users still mostly prefer to write a query using terms that consist of two, one, or three word(s). Again, five groups were created to investigate how many visit interactions were made per session based on the number of visits. The findings are given in Figure 2.

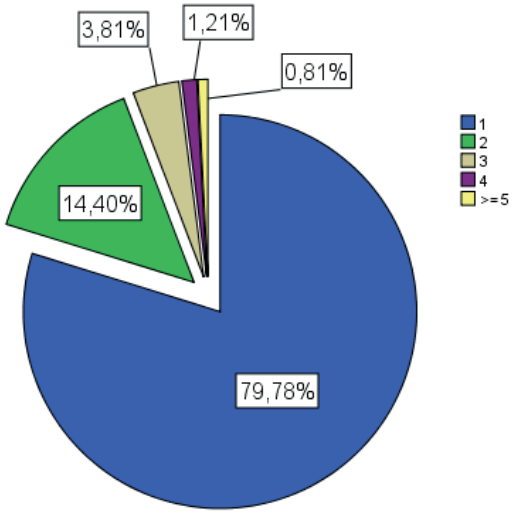

Figure 2. The groups of visit interactions based on the number of visits per session

Even though Jansen and Spink (2005) stated that 30.3% of users examine only one document per session, users' visit interactions in Figure 2 show that almost 80% of users view only one document per session. This can be interpreted as an indication that users have become more impatient while searching. In addition, these findings point out that almost 80% of the TTFC and TTLC values are the same for each session in the prepared datasets. Based on the sessions with no abandonment, the statistics regarding the ranks of documents the users clicked are shared in Figure 3.

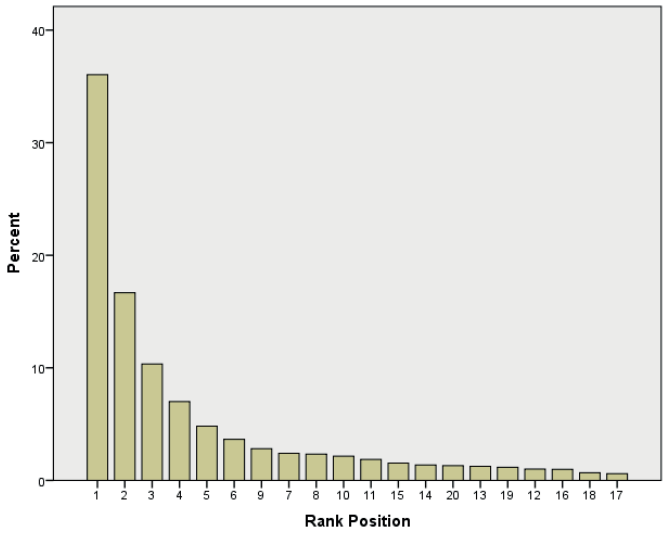

Figure 3. The ranks of the visited documents

Similar to the findings from the studies by Joachims et al. (2005), Agichtein, Brill, Dumais, and Ragno (2006), and Cen et al. (2009), our results show that users mostly prefer to view documents at ranks lower than 10 (Figure 3).

4.2. Descriptive Findings

The descriptive findings that belong to the baseline efficiency indicators for each website group are given in Table 4.

Table 4
The descriptive findings of baseline efficiency indicators

GROUP	A				B				C			
Indicators	**DwSERP**	**TTFC***	**TTLC***	**DwSERP_Ab**	**DwSERP**	**TTFC***	**TTLC***	**DwSERP_Ab**	**DwSERP**	**TTFC***	**TTLC***	**DwSERP_Ab**
Mean	15,52	11,45	60,39	43,24	15,44	12,61	64,87	27,27	14,99	12	60,98	18,03
Median	10,50	9	17	51,10	12,10	11	17	20,20	11,70	10	17	12,20
Std. Dev.	14,38	7,20	144,55	18,12	11,81	7,13	189,66	20,69	11,82	6,87	165,80	16,12
Min.	1	3	3	1	1,50	2	2	1	1	1	1	1
Max.	59,60	28	862	59,60	57,90	30	1785	57,20	59	30	1782	59,90
n		49		1247		199		1118		1899		5311
GROUP	**D**				**E**				**F**			
Indicators	**DwSERP**	**TTFC***	**TTLC***	**DwSERP_Ab**	**DwSERP**	**TTFC***	**TTLC***	**DwSERP_Ab**	**DwSERP**	**TTFC***	**TTLC***	**DwSERP_Ab**
Mean	13,52	11,45	56,86	10,94	11,12	9,73	66,40	11,88	8,95	7,76	129,46	14,97
Median	10,20	10	13	7,45	7,70	8	12	7,90	6,40	6	12	8,80
Std. Dev.	10,92	6,38	138,51	10,28	9,83	6,47	207,12	11,46	8,19	5,23	330,37	15,53
Min.	1	2	2	1	1	1	1	1	1,10	2	2	1
Max.	51,40	29	962	57	49,50	30	1443	59,60	48,70	30	1677	57,20
n		147		304		297		395		139		123

***** These values are in the form of seconds.

In Section 3.2.1, we stated that the website group that managed to keep users most successfully was F, followed by E, D, C, B, and A, respectively. This order also indicates which group performed better than the others. According to the mean values in Table 4, only the DwSERP values carry signs regarding this assumption (here, the mentioned propositions are called "assumptions" because they have not yet been statistically proved). In other words, it can be seen that the mean values decrease from the worst-performing group to the best-performing group. When we focus on the mean TTFC values, this pattern only seems clear for groups D, E, and F. For the mean TTLC values, no consistent pattern was observed. As for the DwSERP_Ab values, only groups A, B, C, and D support the assumptions. Although it is proved in the next section and discussed in Section 5, as a preliminary interpretation, it is better to mention that the DwSERP_Ab indicator shows different characteristics in contrast to the general consensus followed in IR studies. Lastly, as the data of each indicator did not fit the normal distribution (p<0,05), nonparametric test methods were utilized in the analyses. The descriptive findings from the MEDEF-based efficiency indicators are seen in Table 5.

Table 5
The descriptive findings of the MEDEF-based efficiency indicators

GROUP	A							
Indicators	**MRR_DwSERP**	**MRR_TTFC**	**MRR_TTLC**	**MRR_ALL**	**AP_DwSERP**	**AP_TTFC**	**AP_TTLC**	**AP_ALL**
Mean	0,034	0,032	0,021	0,001	0,035	0,035	0,022	0,001
Median	0,020	0,028	0,011	0,00010	0,021	0,028	0,011	0,00010
Std. Dev.	0,038	0,027	0,024	0,004	0,038	0,029	0,024	0,004
Min.	0,00045	0,00101	0,00011	0,0000003	0,00045	0,00101	0,00011	0,0000003
Max.	0,20	0,11	0,08	0,02	0,20	0,11	0,08	0,02
n				49				
GROUP	**B**							
Indicators	**MRR_DwSERP**	**MRR_TTFC**	**MRR_TTLC**	**MRR_ALL**	**AP_DwSERP**	**AP_TTFC**	**AP_TTLC**	**AP_ALL**
Mean	0,046	0,044	0,030	0,001	0,048	0,047	0,031	0,001
Median	0,023	0,028	0,012	0,00008	0,024	0,028	0,013	0,00008
Std. Dev.	0,070	0,051	0,042	0,004	0,071	0,054	0,042	0,004
Min.	0,00086	0,00104	0,00007	0,0000001	0,00115	0,00104	0,00007	0,0000002

Max.	0,53	0,33	0,25	0,04	0,53	0,33	0,25	0,04
n				199				
GROUP				**C**				
Indicators	**MRR_DwSERP**	**MRR_TTFC**	**MRR_TTLC**	**MRR_ALL**	**AP_DwSERP**	**AP_TTFC**	**AP_TTLC**	**AP_ALL**
Mean	0,051	0,044	0,033	0,002	0,054	0,048	0,033	0,002
Median	0,023	0,028	0,013	0,00010	0,026	0,031	0,014	0,00011
Std. Dev.	0,080	0,054	0,052	0,008	0,082	0,056	0,052	0,008
Min.	0,00063	0,00064	0,00003	0,0000001	0,00063	0,00064	0,00003	0,0000001
Max.	0,91	1	1	0,208	0,91	1	1	0,208
n				1899				
GROUP				**D**				
Indicators	**MRR_DwSERP**	**MRR_TTFC**	**MRR_TTLC**	**MRR_ALL**	**AP_DwSERP**	**AP_TTFC**	**AP_TTLC**	**AP_ALL**
Mean	0,053	0,050	0,038	0,002	0,053	0,51	0,038	0,002
Median	0,033	0,033	0,023	0,00020	0,033	0,033	0,023	0,00020
Std. Dev.	0,058	0,051	0,048	0,006	0,060	0,052	0,048	0,006
Min.	0,00123	0,00197	0,00030	0,0000008	0,00178	0,00216	0,00030	0,0000011
Max.	0,36	0,33	0,33	0,042	0,36	0,33	0,33	0,042
n				147				
GROUP				**E**				
Indicators	**MRR_DwSERP**	**MRR_TTFC**	**MRR_TTLC**	**MRR_ALL**	**AP_DwSERP**	**AP_TTFC**	**AP_TTLC**	**AP_ALL**
Mean	0,084	0,070	0,054	0,005	0,085	0,072	0,054	0,005
Median	0,048	0,048	0,031	0,00045	0,050	0,050	0,031	0,00045
Std. Dev.	0,097	0,066	0,066	0,017	0,097	0,067	0,066	0,017
Min.	0,00118	0,00114	0,00010	0,0000003	0,00119	0,00114	0,00010	0,0000003
Max.	0,59	0,5	0,5	0,227	0,59	0,5	0,5	0,227
n				297				
GROUP				**F**				
Indicators	**MRR_DwSERP**	**MRR_TTFC**	**MRR_TTLC**	**MRR_ALL**	**AP_DwSERP**	**AP_TTFC**	**AP_TTLC**	**AP_ALL**
Mean	0,142	0,112	0,076	0,008	0,147	0,119	0,076	0,008
Median	0,077	0,083	0,042	0,00061	0,085	0,089	0,042	0,00067
Std. Dev.	0,173	0,096	0,094	0,025	0,172	0,096	0,094	0,025
Min.	0,00321	0,00529	0,00016	0,0000011	0,00522	0,00529	0,00016	0,0000022
Max.	0,91	0,5	0,5	0,208	0,91	0,5	0,5	0,208
n				139				

With the exception of the MRR_ALL and AP_ALL indicators, a consideration of the mean value of each indicator reveals that all indicators support the assumptions, meaning that the increase in the mean values is directly proportional to the groups' performances. This pattern is not clear for the MRR_ALL and AP_ALL indicators. Even though the TTLC indicator in Table 4 is seen not to be consistent with the assumptions, it was found that the MEDEF positively affected this indicator, making it more consistent, as seen in Table 5. When we focus on each conjugate indicator of each group in Table 5 (such as MRR_DwSERP and AP_DwSERP or MRR_ALL and AP_ALL), it can also be seen that there are slight differences between the indicator values. It is thought that user actions caused this similarity. In other words, as the users mostly viewed only one document in their sessions (Figure 2), the calculation of MRR and AP values was equal for most sessions of these groups. As in the data that belong to the baseline efficiency indicators, the data of the MEDEF-based efficiency indicators did not fit the normal distribution (p<0,05). This is why nonparametric test methods were utilized in the analyses.

4.3. Relationship Among the Indicators

Spearman correlation tests were made on the indicators considering their types. To this end, the findings from the baseline indicators, MEDEF-based indicators created using the MRR metric, and MEDEF-based indicators created using the AP metric are explained, respectively.

Table 6
The findings from the baseline efficiency indicators

INDICATORS	DwSERP	TTFC	TTLC
DwSERP	1,000	,600*	,499*
TTFC	,600*	1,000	,477*
TTLC	,499*	,477*	1,000

* correlation coefficients; $p < 0,01$

According to Table 6, both the DwSERP and TTFC indicators have a positive relationship with the TTLC indicator (r_s: 0,499, 0,477, respectively). As for the relationship between the DwSERP and TTFC indicators, a strong positive relationship was observed (r_s: 0,600, Table 6), which indicates that DwSERP can be foreseen through users' TTFC behaviors and vice versa.

The findings gathered from the MEDEF-based indicators created using the MRR metric are given in Table 7.

Table 7
The findings from the MEDEF-based indicators - MRR

INDICATORS	MRR_DwSERP	MRR_TTFC	MRR_TTLC	MRR_ALL
MRR_DwSERP	1,000	,842*	,714*	,848*
MRR_TTFC	,842*	1,000	,706*	,783*
MRR_TTLC	,714*	,706*	1,000	,905*
MRR_ALL	,848*	,783*	,905*	1,000

* correlation coefficients; $p < 0,01$

Similar but stronger findings were observed for the MEDEF-based indicators as in the baseline indicators (Table 7). In addition, it was revealed that user behaviors can be estimated more consistently from the MEDEF-based indicators based on the MRR metric. These findings also indicate that the MEDEF concretizes the relationship between different user behaviors (the baseline indicators) more clearly. The other findings that belong to the MEDEF-based indicators created using the AP metric are shared in Table 8.

Table 8
The findings from the MEDEF-based indicators - AP

INDICATORS	AP_DwSERP	AP_TTFC	AP_TTLC	AP_ALL
AP_DwSERP	1,000	,829*	,688*	,830*
AP_TTFC	,829*	1,000	,670*	,748*
AP_TTLC	,688*	,670*	1,000	,899*
AP_ALL	,830*	,748*	,899*	1,000

* correlation coefficients; $p < 0,01$

Even though a slight decrease was observed in Table 8 compared to Table 7, our results clearly reveal that the MEDEF-based indicators are more consistent in being able to foresee user behaviors. All findings show the preliminary signs that the MEDEF has a high potential to read user behaviors. The statistical base of whether these signs carry meaning is examined in the next section.

4.4. Investigations of Research Questions

In the scope of the present study, after preprocessing the collected data, twenty-three websites were separated into six groups based on their performances regarding users' Sab behaviors. Whether this grouping is meaningful statistically was investigated with the first research question.

RQ1. What are the differences between the SAb behaviors in the groups?

According to the findings from the Chi-square test made on six groups, the propositions created to guide the other research questions have been found statistically significant (Table 9; $\chi^2 = 638,165$; $p < 0,01$).

Table 9
The performance findings of groups regarding users' SAb behaviors

			Abandonment		Total
			false	true	
Groups	A	Count	49	1247	1296
		% within Groups	3,8%	96,2%	100,0%
	B	Count	199	1118	1317
		% within Groups	15,1%	84,9%	100,0%
	C	Count	1899	5311	7210
		% within Groups	26,3%	73,7%	100,0%
	D	Count	147	304	451
		% within Groups	32,6%	67,4%	100,0%
	E	Count	297	395	692
		% within Groups	42,9%	57,1%	100,0%
	F	Count	139	123	262
		% within Groups	53,1%	46,9%	100,0%
Total		Count	2730	8498	11228
		% within Groups	24,3%	75,7%	100,0%

This research proves that the propositions created presumptively through Table 2 can be followed as guidance to determine the performance of the efficiency indicators. In addition to Table 9, the groups were also compared as pairs (15 comparisons) with the Chi-square test, and it was observed that the results did not change for all comparisons. While answering the other questions, the reflectiveness performances of each efficiency indicator were evaluated through these fifteen performance propositions, and success was determined according to how many propositions could be detected by each indicator.

RQ2. How do baseline efficiency indicators perform in reflecting users' SAb behavior?

Sixty comparisons, 15 for each indicator, were made with the Mann-Whitney U test between the groups to answer this question, and the revealed findings are given in Table 10.

Table 10
The findings regarding the reflectiveness performances of baseline efficiency indicators

Indicators	DwSERP	TTFC	TTLC	DwSERP_Ab
NUS *	9	8	5	12
NUNOS **	6	7	10	3
Rate of NUS	60%	53,33%	33,33%	80%
Rate of NUNOS	40%	46,67%	66,67%	20%

The significance level is chosen as 0,05; $p<0,05$
* The number of comparisons found as significant
** The number of comparisons found as nonsignificant

A consideration of the sessions with no abandonment reveals that the most reliable indicator is DwSERP, followed by TTFC and TTLC, respectively. In addition, strikingly, DwSERP_Ab was determined to be the most successful indicator in reflecting group performances compared to the other baseline indicators (the reason why it is striking is discussed in Section 5).

RQ3. How do MEDEF-based efficiency indicators perform in reflecting users' SAb behavior?

One hundred twenty comparisons, 15 for each MEDEF-based indicator, were made with the Mann-Whitney U test between the groups to answer this question, and the findings are shared in Table 11.

Table 11
The findings regarding the reflectiveness performances of MEDEF-based efficiency indicators

Indicators	MRR_DwSERP	MRR_TTFC	MRR_TTLC	MRR_ALL	AP_DwSERP	AP_TTFC	AP_TTLC	AP_ALL
NUS	12	12	10	10	12	10	11	10
NUNOS	3	3	5	5	3	5	4	5
Rate of NUS	80%	80%	66,67%	66,67%	80%	66,67%	73,33%	66,67%
Rate of NUNOS	20%	20%	33,33%	33,33%	20%	33,33%	26,67%	33,33%

The significance level is chosen as 0,05; $p<0,05$

According to Table 11, DwSERP was found to be the most reliable punishment element regardless of which effectiveness metric is used in the MEDEF. In addition, even though the NUNOS value is three for the MRR_DwSERP and AP_DwSERP indicators, the analyses with the Mann-Whitney U test showed that these indicators tended to reflect the three propositions. For the MRR metric, it is also clear that using TTFC as a punishment element reveals the same results. Moreover, TTLC is in the second position for the AP metric in reflecting group performances successfully. When we consider all MEDEF-based indicators, it can be stated that they have outperformed the reflectiveness performance of the baseline indicators unambivalently.

RQ4. Which prediction model created based on the type of indicators shows more successful performance?

Before creating the prediction models, the data on each dataset were separated into two sets (training and test) depending on the separation rate used. As the purpose was to reach similar findings to the second and third research questions, no preprocessing was applied on all sets, and the data was utilized without a touch. The findings gathered from two different ML algorithms on each dataset are given in Table 12.

Table 12
The performances of created prediction models

Datasets	Algorithms	Separation Rate (Train: Test)	ACC	P	F1 Score
baselines	DT	0,9: 0,1	0,696	0,484	0,571
		0,8: 0,2	0,696	0,484	0,571
		0,7: 0,3	0,659	0,533	0,576
	RF	0,9: 0,1	0,586	0,533	0,554
		0,8: 0,2	0,590	0,542	0,564
		0,7: 0,3	0,602	0,527	0,559
medef_MRR	DT	0,9: 0,1	0,696	0,504	0,578
		0,8: 0,2	0,698	0,512	0,579
		0,7: 0,3	0,698	0,513	0,579
	RF	0,9: 0,1	0,626	0,570	0,592
		0,8: 0,2	0,612	0,565	0,586
		0,7: 0,3	0,624	0,555	0,585
medef_AP	DT	0,9: 0,1	0,700	0,537	0,579
		0,8: 0,2	0,698	0,512	0,579
		0,7: 0,3	0,698	0,513	0,579
	RF	0,9: 0,1	0,623	0,559	0,587
		0,8: 0,2	0,612	0,559	0,582
		0,7: 0,3	0,629	0,552	0,583

When considering the DT algorithm, although the ACC values on each separation rate do not show the performance differences clearly, the P value and F1 score revealed that the MEDEF supported the algorithm in detecting which session belonged to which website group (Table 12). For the RF algorithm, however, the prediction performance of each model was observed more clearly in the process of proving the success of MEDEF, regardless of which type of evaluation metric was considered (Table 12). These findings correspond to the findings in the third research question, meaning that the MEDEF allows user behavior regarding baseline indicators to be interpreted in a more reliable way.

5. DISCUSSION

In the scope of the present study, users' search interactions were recorded through search modules that were integrated into the department websites of a university in Turkey for one month. After preprocessing the collected data, the websites were separated into six groups based on users' SAb behaviors, and fifteen propositions were created regarding the IR performances of these groups. Afterward, 12 efficiency indicators, consisting of four baselines already used in IR studies, and eight created by following the proposed efficiency formula, were compared to each other. While the statistical investigations focused on how many propositions these indicators can reflect and how reliable they are for evaluations and optimizations, the ML-based examinations were carried out to reveal signs that can be interpreted as directly proportional to findings from the statistical investigations.

A consideration of the baseline efficiency indicators reveals that the most reliable indicator was found to be DwSERP for sessions with no abandonment, followed by TTFC and TTLC. In his study, Arguello (2014) found that dwell time is a useful predictor when determining search task difficulty. Even though the author considered the dwell time on landing pages, the findings in the present study are seen to be in the same line, meaning that the dwell time is a reliable indicator regardless of which type of usage is preferred. In the study by Jung, Herlocker, and Webster (2007), the authors tried to improve the quality of search results and emphasized that the last visited documents were useful. Although the context between the present and their study is different, the TTLC indicator was not seen as a reliable efficiency indicator in this study.

A separate parenthesis is required for the DwSERP_Ab indicator. The expectation of DwSERP_Ab is that it will inversely reflect the success of groups in the propositions because the consensus in IR studies is that the decrease in the DwSERP_Ab value points out negative/bad experiences (Song, Shi, White, and Awadallah, 2014; Borisov, Markov, de Rijke, and Serdyukov, 2016). However, with the exception of the E and F groups, Table 4 shows that the increase in group performances and the mean values of the DwSERP_Ab indicator are inversely proportional to each other. These findings are also supported by the analysis results from the Mann-Whitney U test, meaning that the present study shows opposing results. The reason for this could be that after users saw the first response from the search modules, they reformulated their query in a short time and accordingly reached satisfying information in their consecutive queries. The other possibility, as indicated in the study by Stamou and Efthimiadis (2010), could be that users met their information needs by briefly examining only the result snippets on the SERP without spending much time. Regardless of what type of interaction users performed, this result can be interpreted to show that the more "negative abandonment" IR systems have, the more efficient performance users show.

As for the created indicators, it was proved that all eight indicators outperformed the baseline indicators for the sessions with no abandonment. Moreover, the TTFC and TTLC indicators, which performed less consistently than DwSERP, also showed more consistent performance when used as punishments. As DwSERP is already a reliable indicator, the reflectiveness of both the MEDEF-based indicators created with it also showed the most successful performance. In general terms, the MEDEF further strengthened the reflections in baseline efficiency indicators.

6. CONCLUSION

In the present study, a modular efficiency determination formula, MEDEF, is proposed. Using the MEDEF, eight efficiency indicators were created and compared with the baseline indicators already used in IR studies. The findings revealed that the MEDEF-based indicators outperform the baselines. It is believed that indicators created by following the MEDEF will likely show more reliable findings in evaluations and optimizations of IR systems. Moreover, only three different efficiency indicators (the baselines) were used as punishments in this study to create the MEDEF-based indicators. The common part

of these baselines is that they are implicit data, meaning that other implicit data types can also be integrated into the MEDEF. Apart from this, if implicit data, which will be integrated, are a kind of reward rather than a punishment, then the punishment metric can be used reciprocally.

As in the baseline indicators, two effectiveness metrics used in the MEDEF-based indicators were also based on natural user behaviors. In other words, they also are implicit data (clickthrough). For future studies, it is thought that different metrics (such as NDCG or ERR) based on explicit data, which can be gathered from users while they search, can also be integrated into the MEDEF.

In conclusion, the present study revealed the trustworthiness and consistency of four different efficiency indicators already in use in IR studies for the purpose of optimization. In addition, it was determined that the MEDEF-based indicators were more successful than these indicators. This result can be interpreted to show that the indicators created by following the MEDEF will boost the success of optimization-based IR studies in the future. Moreover, the question of how IR systems can be evaluated using only efficiency indicators was enlightened. While the reflectiveness performance of baseline efficiency indicators (except for DwSERP_Ab) is not adequate for individual usage, the MEDEF-based indicators showed successful performance, meaning that the MEDEF-based indicators can be utilized individually to evaluate IR systems from the usability perspective, as exemplified in the present study.

Peer-review: Externally peer-reviewed.
Conflict of Interest: The author has no conflict of interest to declare.
Grant Support: The author declared that this study has received no financial support.

REFERENCES

Agichtein, E., Brill, E., & Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, 19–26. New York, New York, USA: ACM Press. https://doi.org/10.1145/1148170.1148177

Agichtein, E., Brill, E., Dumais, S., & Ragno, R. (2006). Learning user interaction models for predicting web search result preferences. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '06*, 3. New York, New York, USA: ACM Press. https://doi.org/10.1145/1148170.1148175

Alhabashneh, O., Iqbal, R., Doctor, F., & James, A. (2017). Fuzzy rule based profiling approach for enterprise information seeking and retrieval. *Information Sciences*, 394–395, 18–37. https://doi.org/10.1016/J.INS.2016.12.040

Arguello, J. (2014). Predicting search task difficulty. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 8416 LNCS (pp. 88–99). https://doi.org/10.1007/978-3-319-06028-6_8

Arkhipova, O., Grauer, L., Kuralenok, I., & Serdyukov, P. (2015). Search Engine Evaluation based on Search Engine Switching Prediction. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 723–726. New York, NY, USA: ACM. https://doi.org/10.1145/2766462.2767786

Aula, A., Khan, R. M., & Guan, Z. (2010). How does search behavior change as search becomes more difficult? *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*, 35. New York, New York, USA: ACM Press. https://doi.org/10.1145/1753326.1753333

Balakrishnan, V., & Zhang, X. (2014). Implicit user behaviours to improve post-retrieval document relevancy. *Computers in Human Behavior*, 33, 104–112. https://doi.org/10.1016/J.CHB.2014.01.001

Beierle, F., Aizawa, A., Collins, A., & Beel, J. (2020). Choice overload and recommendation effectiveness in related-article recommendations. *International Journal on Digital Libraries*, 21(3), 231–246. https://doi.org/10.1007/s00799-019-00270-7

Belkin, N. J., Kelly, D. F., Kim, G., Kim, J. Y., Lee, H., Muresan, G., Tang, M. C., Yuan, X., Cool, C. (2003). Query Length in Interactive Information Retrieval. SIGIR Forum (ACM Special Interest Group on Information Retrieval), (SPEC. ISS.), 205–212. New York, New York, USA: ACM Press. https://doi.org/10.1145/860472.860474

Borisov, A., Markov, I., de Rijke, M., & Serdyukov, P. (2016). A Context-aware Time Model for Web Search. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 205–214. New York, NY, USA: ACM. https://doi.org/10.1145/2911451.2911504

Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval - SIGIR '04*, 25. New York, New York, USA: ACM Press. https://doi.org/10.1145/1008992.1009000

Büttcher, S., Clarke, C. L. A., & Cormack, G. V. (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. The MIT Press.

Cen, R., Liu, Y., Zhang, M., Ru, L., & Ma, S. (2009). Study on the click context of web search users for reliability analysis. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 5839 LNCS (pp. 397–408). https://

doi.org/10.1007/978-3-642-04769-5_35

Chapelle, O., Metlzer, D., Zhang, Y., & Grinspan, P. (2009). Expected reciprocal rank for graded relevance. *Proceeding of the 18th ACM Conference on Information and Knowledge Management - CIKM '09*, 621. New York, New York, USA: ACM Press. https://doi.org/10.1145/1645953.1646033

Claypool, M., Brown, D., Le, P., & Waseda, M. (2001). Inferring user interest. *IEEE Internet Computing*, 5(6), 32–39. https://doi.org/10.1109/4236.968829

Colace, F., De Santo, M., Greco, L., & Napoletano, P. (2015). Improving relevance feedback-based query expansion by the use of a weighted word pairs approach. *Journal of the Association for Information Science and Technology*, 66(11), 2223–2234. https://doi.org/10.1002/asi.23331

Croft, B., Metzler, D., & Strohman, T. (2009). *Search Engines: Information Retrieval in Practice* (1st ed.). Boston: Pearson.

Diriye, A., White, R., Buscher, G., & Dumais, S. (2012). Leaving so soon?: understanding and predicting web search abandonment rationales. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management - CIKM '12*, 1025. New York, New York, USA: ACM Press. https://doi.org/10.1145/2396761.2398399

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. Retrieved from http://scikit-learn.sourceforge.net.

Fox, S., Karnawat, K., Mydland, M., Dumais, S., & White, T. (2005). Evaluating implicit measures to improve Web search. *ACM Transactions on Information Systems*, 23(2), 147–168. https://doi.org/10.1145/1059981.1059982

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '00*, 345–352. New York, New York, USA: ACM Press. https://doi.org/10.1145/332040.332455

Hassan, A. (2012). A semi-supervised approach to modeling web search satisfaction. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '12*, 275. New York, New York, USA: ACM Press. https://doi.org/10.1145/2348283.2348323

Hassan, A., Song, Y., & He, L. (2011). A task level metric for measuring web search satisfaction and its application on improving relevance estimation. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management - CIKM '11*, 125. New York, New York, USA: ACM Press. https://doi.org/10.1145/2063576.2063599

Hassan, A., White, R. W., Dumais, S. T., & Wang, Y.-M. (2014). Struggling or exploring? Disambiguating Long Search Sessions. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, 53–62. New York, NY, USA: ACM. https://doi.org/10.1145/2556195.2556221

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, 64(2), 79–102. https://doi.org/10.1016/J.IJHCS.2005.06.002

Jansen, B. J., & Spink, A. (2005). An analysis of Web searching by European AlltheWeb.com users. *Information Processing & Management*, 41(2), 361–381. https://doi.org/10.1016/S0306-4573(03)00067-0

Jansen, B. J., Spink, A., Bateman, J., & Saracevic, T. (1998). Real life information retrieval: a study of user queries on the Web. ACM SIGIR Forum, 32(1), 5–17. https://doi.org/10.1145/281250.281253

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management*, 36(2), 207–227. https://doi.org/10.1016/S0306-4573(99)00056-4

Jarvelin, K., & Kekalainen, J. (2000). IR evaluation methods for retrieving highly relevant documents. SIGIR Forum (ACM Special Interest Group on Information Retrieval), 41–48. New York, New York, USA: ACM Press. https://doi.org/10.1145/3130348.3130374

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4), 422–446. https://doi.org/10.1145/582415.582418

Jiang, D., Leung, K. W. T., Yang, L., & Ng, W. (2015). Query suggestion with diversification and personalization. *Knowledge-Based Systems*, 89, 553–568. https://doi.org/10.1016/J.KNOSYS.2015.09.003

Joachims, T. (2002). Optimizing search engines using clickthrough data. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 133–142. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/775047.775067

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. *SIGIR 2005 - Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 154–161. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1076034.1076063

Jung, S., Herlocker, J. L., & Webster, J. (2007). Click data as implicit relevance feedback in web search. *Information Processing & Management*, 43(3), 791–807. https://doi.org/10.1016/J.IPM.2006.07.021

Kelly, D., & Fu, X. (2007). Eliciting better information need descriptions from users of information search systems. *Information Processing & Management*, 43(1), 30–46. https://doi.org/10.1016/J.IPM.2006.03.006

Kim, Y., Hassan, A., White, R. W., & Zitouni, I. (2014). Modeling dwell time to predict click-level satisfaction. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, 193–202. New York, NY, USA: ACM. https://doi.org/10.1145/2556195.2556220

Kuhar, M., & Merčun, T. (2022). Exploring user experience in digital libraries through questionnaire and eye-tracking data. *Library & Information Science Research*, 44(3), 101175. https://doi.org/10.1016/J.LISR.2022.101175

Lee, C.-J., Teevan, J., & de la Chica, S. (2014). Characterizing multi-click search behavior and the risks and opportunities of changing results during use. *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 515–524. New York, NY, USA: ACM. https://doi.org/10.1145/2600428.2609588

Liu, Y., Chen, Y., Tang, J., Sun, J., Zhang, M., Ma, S., & Zhu, X. (2015). Different Users, Different Opinions. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 493–502. New York, NY, USA: ACM. https://doi.org/10.1145/2766462.2767721

Liu, C., Gwizdka, J., & Liu, J. (2010). Helping identify when users find useful documents: examination of query reformulation intervals. *Proceeding of the Third Symposium on Information Interaction in Context - IIiX '10*, 215. New York, New York, USA: ACM Press. https://doi.org/10.1145/1840784.1840816

Liu, C., Liu, J., & Belkin, N. J. (2014). Predicting Search Task Difficulty at Different Search Stages. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 569–578. New York, NY, USA: ACM. https://doi.org/10.1145/2661829.2661939

Makkar, A., & Kumar, N. (2018). User behavior analysis-based smart energy management for webpage ranking: Learning automata-based solution. *Sustainable Computing: Informatics and Systems*, 20, 174–191. https://doi.org/10.1016/J.SUSCOM.2018.02.003

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. New York: Cambridge University Press.

Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1), 1–27. https://doi.org/10.1145/1416950.1416952

Nasir, J. A., Varlamis, I., & Ishfaq, S. (2019). A knowledge-based semantic framework for query expansion. *Information Processing & Management*, 56(5), 1605–1617. https://doi.org/10.1016/j.ipm.2019.04.007

Nie, L., Jiang, H., Ren, Z., Sun, Z., & Li, X. (2016). Query Expansion Based on Crowd Knowledge for Code Search. *IEEE Transactions on Services Computing*, 9(5), 771–783. https://doi.org/10.1109/TSC.2016.2560165

Nielsen, J. (1993). *Usability Engineering* (1st ed.). Morgan Kaufmann.

Radlinski, F., Kurup, M., & Joachims, T. (2008). How does clickthrough data reflect retrieval quality? *Proceeding of the 17th ACM Conference on Information and Knowledge Mining - CIKM '08*, 43. New York, New York, USA: ACM Press. https://doi.org/10.1145/1458082.1458092

Rosenzweig, E. (2015). *Successful User Experience: Strategies and Roadmaps* (1st ed.). Morgan Kaufmann.

Rubin, J., & Chisnell, D. (2008). *Handbook Of Usability Testing How To Plan Design And Conduct Effective Tests*. Wiley Pub.

Sarkar, S., Mitsui, M., Liu, J., & Shah, C. (2020). Implicit information need as explicit problems, help, and behavioral signals. *Information Processing & Management*, 57(2), 102069. https://doi.org/10.1016/J.IPM.2019.102069

Singh, J., & Sharan, A. (2017). A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach. *Neural Computing and Applications*, 28(9), 2557–2580. https://doi.org/10.1007/s00521-016-2207-x

Singla, A., & White, R. W. (2010). Sampling high-quality clicks from noisy click data. *Proceedings of the 19th International Conference on World Wide Web - WWW '10*, 1187. New York, New York, USA: ACM Press. https://doi.org/10.1145/1772690.1772867

Song, Y., Shi, X., White, R., & Awadallah, A. H. (2014). Context-aware web search abandonment prediction. *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 93–102. New York, NY, USA: ACM. https://doi.org/10.1145/2600428.2609604

Stamou, S., & Efthimiadis, E. N. (2010). Interpreting user inactivity on search results. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 5993 LNCS (pp. 100–113). https://doi.org/10.1007/978-3-642-12275-0_12

Vidinli, I. B., & Ozcan, R. (2016). New query suggestion framework and algorithms: A case study for an educational search engine. *Information Processing and Management*, 52(5), 733–752. https://doi.org/10.1016/j.ipm.2016.02.001

Voorhees, E. M. (1999). The TREC-8 Question Answering Track Report. *Text REtrieval Conference*, 77–82.

Zhai, C., & Massung, S. (2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. ACM. https://doi.org/10.1145/2915031