Parsisanj: an automatic component-based approach towards search engine evaluation.*

Amin Heydari Alashti[1][0000−0002−3689−5796], Ahmad Asgharian Rezaei[2][0000−0002−8510−8875], Alireza Elahi[3], Mohammad Ali Abam[4], and Mohammad Ghodsi[4]

[1] Snapp, Iran
amin.heydarialashti@snapp.cab
[2] RMIT University
ahmad.asgharian.rezaei@rmit.edu.au
[3] Shahid Beheshti University, Tehran, Iran
[4] Computer Science Facutly, Sharif University of Technology, Tehran, Iran

Abstract. Web search engines play a significant role in answering users' information needs based on the huge amount of data on the internet. Although evaluating the performance of these systems is too urgent for their improvement, there is no comprehensive, unbiased, low-cost, and reusable method for this problem. Previous works used a small and limited set of queries for their evaluation process that restricts the assessment domain. Moreover, manually search engines evaluation makes the analysis's result subjective; consequently, it would be a high cost and a hard to redo evaluation. Related works considered search engines as a block-box system, and it made their evaluation to be focused just on the ranker component of search engines. In this research, we propose an automatic approach for search engine evaluation that is based on the structure of their components. Thus, it enables us to have a detailed analysis of each of the search engine's components' expertise level. Experimental results of applying thousands of queries on two Persian and two other language-independent search engines show that the two latter engines beat the two former ones in most of the cases; however, the two Persian engines make an acceptable level of proficiency in language-specific sub-systems.

Keywords: Automatic Search Engine Evaluation · Component-based Search Engine Evaluation · Yooz · Parsijoo · Google · Bing.

1 Introduction

Nowadays, the internet has become a prevalent tool that people utilize to answer their information needs. By and by, the information generators grow, so the amount of available data on the web increases similarly. When the amount and variety of provided data on the web increases, the process of finding the exact

* Supported by Iran Telecommunication Research Center.

best match web page to an information need will be much complicated in different aspects. The best match can vary from time to time because a new website's launch can beat some other older websites in some domains. The search engine as a tool that enables people to find the web page, which answers their needs in the best way at the moment, solves the problem of the huge amount of data on the web.

The large number of websites make the testing process of search engines more challenging, and the testing process is urgent for each system to finds its weaknesses. The challenges are high time and price cost, its results' reliability, repeatability, coverage, etc. These are among the most interactive goals that attract researchers to improve. Moreover, other aims that some conducted studies followed in evaluating search engines' different aspects like finding the best search engine for a specific domain.

Search engines have been evaluated mostly by small sets of queries in terms of number and category, manual assessment, and partial evaluation. All the previously conducted researches in this domain struggle with one or more of the mentioned problems: the test's query set was too small or just contained some specific domain of the information a general search engine covers, the query-set contained just one type of query e.g. navigational queries, the main evaluation section of the assessment ran manually that is high cost and also subjective, and they all just evaluated the ranker component of the search engines.

Each search engine regularly consists of multiple components like query analyzer, web crawler, document indexer, and ranker. These are the greatest components of a search engine, but from the user's point of view, the query analyzer's subsystem's and ranker's performance have a direct and great impact on the result of users' queries. If an input query of a search engine is considered as a document as like as its crawling web pages, each input document should be parsed and preprocessed before any analytical process can be applied to it. The analytical parsing and preprocessing modules are located at the query analyzer component; therefore, any existing bug in this preprocessing step can change the result of the input queries. Furthermore, the ranker component utilizes various features to sort the analogous-detected pages, and if mistakenly some features are considered or neglected, its side-effects will impact the input queries' results. That being said, there should be a test evaluation to assess these components in the search engine development process to improve the system continuously.

In the present work, an automatic component-based search engine evaluation method is proposed. According to our knowledge, it is a breakthrough contribution in comparison with all previous similar systems in terms of the number of queries, query types coverage, evaluation methods, low cost, repeatability, result consistency, and reliability.

This paper is organized as follows. Related work is introduced and discussed in Section 2. Our in section 3. Our method will be discussed in Section 3. In Section 4, the assessment results will be presented. Conclusions and future work are presented in Section 5.

2 Related Work

Previous works can be categorized based on different features like the structure of the query-set, the query-set size, the evaluation type, and the study's generality level. In the following paragraphs, related works will be discussed based on their features. The query-set structure is modified using various approaches like putting a constraint on using specific query types, different construction methodologies, and their query sources. Some like [1] constrained their query-set to navigational queries to compare the performance of search engines. Some other studies used a mixed set of informational and navigational queries like [5]. IR datasets like TREC are one of the query sources that some researchers utilized for extracting their query set.[7] But the most prevalent method of building query-set in related works is selecting a set of keywords by crowd-sourcing and extracting keywords from available documents like academic papers' keywords, search engines' search log files. [1–6, 8–10]

Most of the studies use a small set of queries to decline the cost of the manual evaluation. It turns out that the type of the query is an important factor that impacts the size of the query-set. Accordingly, the only work that has an acceptable size is [1] which contains 2000 queries, but all the queries are navigational. The next greatest query-set size used in [3], 400 queries, that is based on crowd-sourcing and a selected subset of search engines' query log; however, usually search engine's query log is not an accessible source for anyone.[7] A query-set size of 200 is the next largest set that is extracted from TREC dataset. Other studies built keyword-based datasets using crowd-sourcing or extracting paper's keywords. Therefore, to docile the cost, their query-set size was too small. [3, 4, 6–9]

As a matter of fact, due to the large and diverse amount of data gathered and processed in search engines and the complex architecture of components it contains, it would be impossible to evaluate their behaviour with just a small set of queries. Previous works have assessed a small portion of search engines' components which cannot give a complete illustration of their weak points and strengths. Thus, to achieve the goal of building a general roadmap for improving search engines, a well-defined query-set with an adequate number of queries in number and domain is needed.

The employed evaluation type in a study is again highly related to the type of queries in its query-set. [1] has provided a set of 2000 Persian navigational queries and submit them to Google, Bing, and Parsijoo. Although it had the largest query-set, due to the nature of navigational queries that have just a single correct answer, the assessment could be done automatically. Of course, the assessment's aim and available meta-data are the other evaluation type selectors' factors. For instance, [4] assessed the overlap and coverage of search engines, so it could test its studying search engines' results by aggregating and comparing the returning links of search engines automatically. On the other hand, [5] uses some computational linguistics datasets to evaluate the hit count returned by search engines for each query. But it used a manual approach to test if its automatic evaluation method is reliable or not.

As mentioned before, due to the small set of queries the previous related studies used, they could only test a small portion of the domains that a search engine covers. They have just tested the ambiguity level in queries that search engines can handle [7]; the coverage of special-purpose websites, like national language websites and the websites that are active in a special science category, by considering navigational queries [1, 8]; finding a search engine that returns the most robust *hit count* property for computational linguistics' research domains [5].
To sum up, here are the common problems in related works' evaluation methods:

- using a small set of features cannot illustrate a real view of a large and multi-aspect system like a search engine.
- Manual assessment will bring subjectivity in evaluation results.
- regardless of appraising a search engine according to users' view-point, ranking is not the only component that should be assessed. Actually, there are also other components that will impact users' experience that should be evaluated.

This study's main aim is to eliminate all the limitations that challenged related works. It tries to defeat the small size of the query-set; meanwhile, an automatic evaluation method is proposed. A structure-sensitive approach surrounded all the methods and approaches in this research that made the break-through difference between this work and the previous ones and is discussed in the following section. This work tries to make a novel component-based evaluation method on search engines. The workflow of designing such a system is as follow:

1. Identifying components of a search engine.
2. designing evaluation domains that are the backbone of evaluating each component with multiple difficulty levels.
3. designing metrics for calculating relevancy score of search engine's result for a query.
4. depicting the step-by-step roadmap by which the query-set should be designed.

Each of the above steps is built on the gained knowledge in its previous step. Generally, by finding a detailed picture of a search engine's architecture, we built an assessment roadmap for each component. A detailed designed query is submitted, then checks if the specified features of the query exist in the returned results of the search engine or not.

3 Method

3.1 Components

Each input of a search engine should be transformed to a unified form of data using some pre-processing steps as a part of a component that is usually called query analyser. The input can be a fetched webpage or users' queries. This unifying step prepares input data for further processes and improves performance

accuracy of other components of a search engine.
Evaluation domains define the type and specifications of queries for each of the components precisely. This is one of the main contributions of this work that covers most of the critical challenges the components face within production environments. Moreover, it has a hierarchy of difficulty that enables to measure search engines' level of expertise in each component.
In the following part, each sub-system will be introduced and its evaluation domains will be discussed.

Normalizer is responsible to add or remove some parts of the input text to modify it to a common form between all the input types of the search engine using a well-defined mapping function. In Persian text, some similar Persian and Arabic characters can be used interchangeably in some words that makes different spellings of a single term. Besides, about 70% of Persian alphabets have similar shapes and sounds. As a result, the multi-shaped words with a single meaning or multiple meanings are an important issue in Persian text pre-processing. To recap, this sub-system should map multi-shaped terms to a single shape to increase text-matching accuracy.
Evaluation domains to assess search engine's ability to normalize the inputs are:

1. Mapping numbers to written form and vice-versa
   1.1 Cardinal numbers
   1.2 Ordinal numbers
   1.3 Cost and benefits
   1.4 Time
   1.5 Date
   1.6 Population
2. Single words with multiple written forms
   2.1 Hamzeh based multi-form words
   2.2 Character repetitions with similar sounds
   2.3 Detecting correct character's initial, centric and final forms
3. Words with a single sound but different written forms
   3.1 All are live words
   3.2 Just one form is live

Tokenizer is responsible for splitting an input text to its meaningful finer granularities like paragraphs, sentences, phrases, and terms. In a search engine, finding these smaller parts is a key point to build an efficient and effective index and ranking process. Persian has eight basic verb structure and multiple types of compound verbs.[11] Compound lingual structures are the most difficult parts of the language to tokenize. Verb, noun, and adjective phrases are built using some language features. E.g. Ezafe forms noun phrases, but the problem is that it is not written and is just pronounced, so its recognition is a challenging problem. Therefore, processing Persian text needs a robust tokenizer to solve its large number of challenges. Evaluation domains to assess search engine's ability to tokenize the inputs are:

1. terms are joint without separator
2. phrase detection
   2.1 Two-parted verbs
   2.2 Three- to five-parted verbs that at least one of them has plural suffix
   2.3 Named entities prepended by identifiers

Spell Correction checks if any term in the input text is out of language's vocabulary, or if a term does not match the context. If finds any typo, it will suggest the most probable candidates or substitutes it with the most probable candidate. Input text of search engines from both directions(user input, fetched webpages) may have typos. Having typos means increasing false negatives in the matching process. So, employing robust solutions can improve candidate documents matching for users' queries. Evaluation domains to assess search engine's ability to find and correct typos in the inputs are:

1. Lexicon
2. Inflection
3. Homonyms
4. Frequency of words
5. Keyboard order

Query expansion is responsible for moving a general context query to a more specific context by adding additional information. Firstly, it was just based on some ontologies like wordnet and language models that were built on the web data. Secondly, it moves on using gathered information from users and helped search engines to propose a personalized search result. Adding some terms and phrases to a query to make it more specific is too risky that can result in increased false positive. Evaluation domains to assess search engine's ability to expand queries are:

1. Synonyms
2. Abbreviations
3. Punctuations

Ranking Its output is a ranked list of extracted candidate documents that are relevant to the input query. Errors of previous steps can spread to this step, but it is premised that previous steps of the evaluating component are errorless. When a query is sent to a search engine, the query analyser will process it and finds user's intention represented in it. The processed query is used to select a candidate list of pages to be the input of the rank component. Subsequently, the rank component uses a complex score function to rank the candidates based on the query. All the previous works concentrated on evaluating this component manually, so their assessment is subjective. They might not considered the fact that the automatic evaluation of the rank component does not necessarily mean implementing its complex score function which is impossible and unfair to implement in the assessment process. Evaluation domains to assess search engine's ability to rank the relevant candidates are:

1. Navigational Queries
2. Trends with single URL
3. Known items

3.2 Evaluation Features

Features help to build an automatic and highly precise query evaluation system. The rudimentary webpage relevancy check is searching for the query's extracted features on its content. Combining these features makes the total structure of this system's score functions. Features of various types can be divided into the below categories:

1. Content-based: occurrence of metrics and their frequency, content's length, ...
2. Structure-based: cares about the occurrence of content-based metrics in different parts of a webpage disparately.
3. Based on result sets structure: inverse document frequency(IDF), and mean reciprocal ranking(MRR) are from this type which considers a result set's structure to evaluate a page's score.
4. Webpage's domain-based: authority and hub domains are always more trustful than regular ones.
5. Hybrid: a combination of the above categories can result in a general evaluation scenario for each component. Content-based and structure-based metrics can be divided into this type.

Moreover, the features' value is calculated based on their type:

1. Shallow Features: have a concrete calculation function E.g. publish time, the occurrence of a specific script in page content, age of the host, Alexa rank, etc.
2. Inference-based features: need some information from the *universal set*[5]. The universal set[6] members are selected based on their close relationship with a query's relevant results. Thus, They can help to distinguish relevant results of a query based on inference-based features. Decision network, which is discussed in following parts, combines these features effectively in calculating relevancy score of a webpage. Some examples of these features are the occurrence of descriptive terms, the occurrence of exclusive terms, document length, URL depth, document readability, etc.

3.3 Query-set Construction

Queries' structure is one of the most distinctive aspects of Parsisanj in comparison with the related works. A Query is precisely well-designed to evaluate

[5] In this system, the universal set contains a set of related webpages to a query. They are used to elicit a value interval for query features by which the relevancy score of a webpage can be calculated in terms of each of its features.

[6] The U set

a certain component of a search engine; furthermore, they contain complementary information to let the system distinguish the relevant answer of the query. Moreover, designers of the query-set might have specified some documents to some of the queries as their universal set. All the above-mentioned sections of a designed query helps to assess the level of expertise of components of search engines. It provides the query's body, the features to identify the relevant and irrelevant answers, and the data to calculate the score of a result webpage.

3.4 Score Functions

Score functions are usually implementing evaluation features and assess a result webpage to find whether it is relevant to the query's topic and information need or not. Designing score function is a challenging part of this research. Some features like shallow features have an explicit mathematical function, and their value is calculated based on result webpages of the query. However, the hybrid features' value of a webpage is calculated using decision network, which is described in following paragraphs.
The similarity of document D and query Q is calculated using:

$$S_{\Lambda}(D;Q) = \Sigma_j \lambda_j . f_j(D,Q) \tag{1}$$

In equation 1, $f_j(D,Q)$ is function j that evaluates document D according to query Q. λ_j is the importance coefficient of feature j; and Λ is the collection of all the λs. Λ is calculated using:

$$argmax_{\Lambda} E(R_{\Lambda};T) \tag{2}$$

In equation 2, $E(R_{\Lambda};T)$ is the assessment parameter that is defined between the score of results given by Parsisanj(R_{Λ}) and the scores given by an expert(T). It shows how much Parsisanj's assessment is analogous to experts' assessment. In other words, features' coefficients Λ should be tuned to converge the score function's output to experts' assessments on a test-set. The size of the test-set is estimated using *hypothesis testing* to ensure that the assessment parameters can represent an ideal set of Λ. About 4 percent of the query-set is found a reasonable size to show that the Parsisanj's and experts' evaluation results are similar ($p<0.05$, independent two-tailed).

3.5 Decision Network

Various factors affect a feature's score in a webpage; and the most challenging type of features are the hybrid ones. Thus, to handle the complexity of hybrid features, decision networks are employed. A decision network is a directed acyclic weighted network, and a node can represent the occurrence of a phrase in content or specifically in one of the assessing webpage sub-parts [12]

In figure 1, Θs are external dependencies that are calculated according to the related documents in the U set of a query. As an example, in searching a

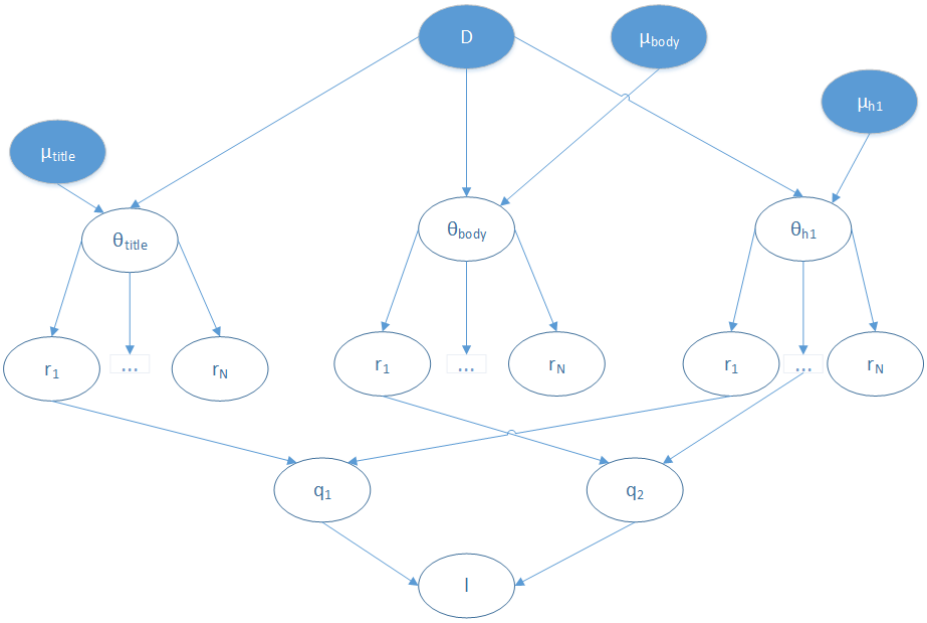


Fig. 1: Decision network of a document D that its relevancy to the queries q_1 and q_2 is evaluated using metrics r_1,r_2,...,r_n.

descriptive term in a document, if the number of its occurrences on related documents of the U set is between 10-15, then it is anticipated to have such a similar behaviour in other relevant documents too. The universal set helps to find a correct distribution of descriptive and exclusive terms in the relevant documents that in terminology of decision network is called Θ.
The root of the decision network is the evaluating document; and its children are different parts of the document. Each part contains its related part of the U set. For instance, the title part of the document contains the title of the U set pages. Moreover, nodes of each part of the document(Θ_{body}) are connected to the features(r_i) that can be defined in that part.
The method for calculating the value of the features in the network is straightforward based on their definitions, and each part of the document's importance is features' value that is normalized by summation of the all features' value for the U set pages. E.g. the following formulas calculate the value(r) and weight(w) of the random variable correspondence to occurrence frequency of a descriptive term in the title of a page:

$$r = count(m; d) \tag{3}$$

$$w = count(m; d)/\Sigma_{u \in U} count(m; u) \tag{4}$$

But to calculate features' value that need inference networks, underlying operators [12, 13] are employed:

$$Belief_{wand}(q_i) = \Pi r_i w_i \tag{5}$$

$$Belief_{wsum} = (\Sigma w_i r_i)/\Sigma w_i \tag{6}$$

Weight of the arcs in an inference network are called belief [12]. Equation 5 is used to calculate score of a feature, and equation 6 is used to calculate the total score of a document that is returned as the result of a query. Thus, equation 1 turns into:

$$f_1 = (\Sigma_{i=1..n} w_{e_i} (\Pi_{j=1} r_{ij} w_j)/\Sigma_{i=1..n} w_{e_i} \tag{7}$$

$$S_\Lambda(D;Q) = \lambda_1.f_1(D,Q) + \lambda_2.f_2(D,Q) + ... + \lambda_k.f_k(D,Q) \tag{8}$$

Equation 7 is used to calculate the features that need the inference network for calculating their scores. To accumulate the total score of a query on a search engine:

$$Score(D_s;Q) = \Sigma_i(S_\Lambda(D_i;Q)/\log 2_i), D_s = \{D_1, D_2, ..., D_m\} \tag{9}$$

The most similar part of Parsisanj to a real search engine is its score function. It might be a suspicion that Parsisanj tries to implement the score function of a search engine; hence, its ranking is not fair and reliable. But there are key differences between what Parsisanj does and what search engines do; which makes the declared assumption false. A list of differences is presented in Table 1.

Table 1: The key differences between Parsisanj's score function and search engines' ranking component.

title	**Parsisanj**	**Search Engine**
involving features	predefined per query	extracting on the fly based on query's information need
#processing pages	a limited number of the top results of each query	scan its whole index for each query

Firstly, table 1 shows that there is no need to be a search engine to have a great ranking algorithm; In other words, Parsisanj moved manually evaluation of search engines from evaluating the result pages of a set of query to the very first step of designing the query set. Therefore, it can evaluate search results of a large array of search engines without any further cost, and simultaneously diminishes the risk of subjectivity in evaluating search engines.
Secondly, the small amount of evaluating result pages helps to be much faster in the relevancy check process, and consequently, it can test various score functions to improve its evaluation process.

4 Results

Hereafter, we will discuss the search engines evaluation outcomes in two phases; each of these two contains some more fine-grained steps in which we will illustrate weak-points and strengths of search engines.
Query Analyzer's modules evaluation part evaluates about 63 thousand result pages of queries. It includes evaluating Normalizer, Tokenizer, Query expansion, and Spell-Checker modules of search engines.

4.1 Normalizer

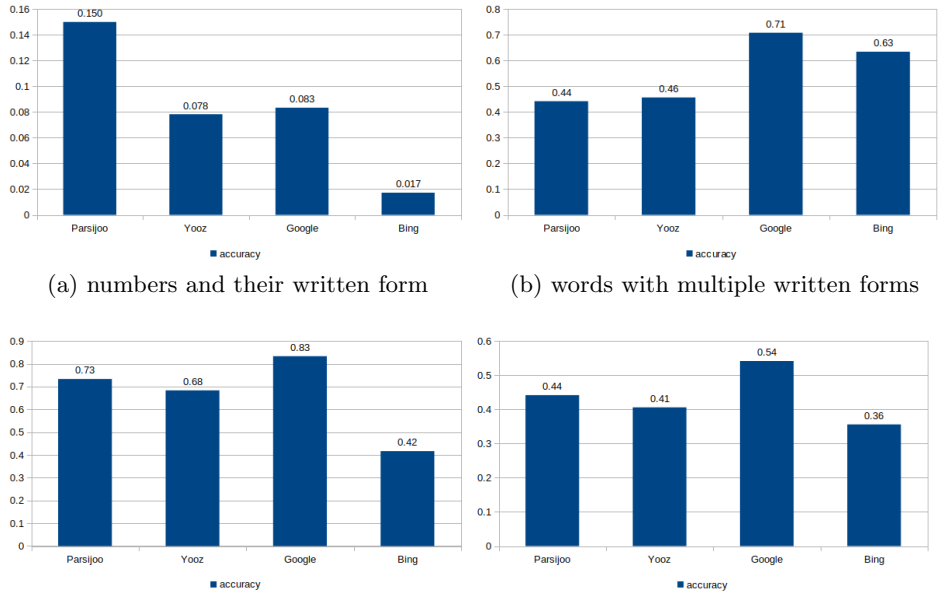
By and large, it was believed that the performance of most of the sub-tasks in the text normalization step has a direct relation to the amount of language-specific knowledge is utilized in designing the systems.

Conversion between numbers and their written form figure 2:a shows that all the search engines have a fundamental problem in converting numbers to their written form and vice-versa. However, supporting the normalization step of all the languages is not expected from international search engines, a coverage of less than eight percent is too disappointing. Regardless of the international search engines' results, it is obvious that the two national ones have not cared about this step. Consequently, it can be the source of further performance problems in other downstream tasks like ranking.

Words with multiple written forms figure 2:b shows that Google and Bing can find different written forms of words regardless of being a multi-lingual international search engine. Yooz and Parsijoo are close in this step, but they have not achieved Google's and Bing's performance. Moreover, it can confirm that utilizing a huge amount of data beside a robust statistical method can result in an excellent level of performance in this normalization subtask.

Homophones Bing in comparison with the other search engines have a much lower level of expertise in handling Persian homophones (Figure 4). However, the national search engines results are much lower than Google's result which might have not put any language-specific knowledge in these modules.

To sum up, in figure 2:d it is obvious that Google has made a better normalization pipeline for Persian contents. The second best search engine is Parsijoo; by regarding its much lower index size, it made a great job in comparison with Google's normalization score. Additionally, with a mere difference in score, Yooz follows Parsijoo. Placed in the fourth level, Bing could not achieve an acceptable performance in the normalization pipeline. Eventually, Google's results reject our first assumption about needed language-specific background knowledge for addressing normalization tasks.



(a) numbers and their written form

(b) words with multiple written forms

(c) homophones

(d) sum up

Fig. 2: Normalizer evaluation

4.2 Tokenizer

This part's queries consist of distinguishing concatenated words and detecting multi-word verbs. In figure 3, the performance of the search engines is presented. Yooz as the best and Google as the second-best search engine solved this challenge. The source of Yooz's excellent performance might be a language-specific approach in tokenization process that beats's Google's method. Furthermore, none of the engines could reach an acceptable performance in detecting multi-word verbs, so the results of this type of questions are not published.

4.3 Spell Correction

figure 4 shows the accuracy of spell correction module of search engines. Parsijoo has designed a much more robust spell correction system rather than Yooz and Bing. It might have achieved a better result even better than Google if it had a huge amount of data that Google Utilizes. Bing shows that besides not having language specific considerations for non-English spell correction, its design lacks the ability to address such tasks using statistical methods.

4.4 Query Expansion

Expanding query with synonyms set in this task performance of a system can be easily influenced by the amount of indexed data in the search engine. So

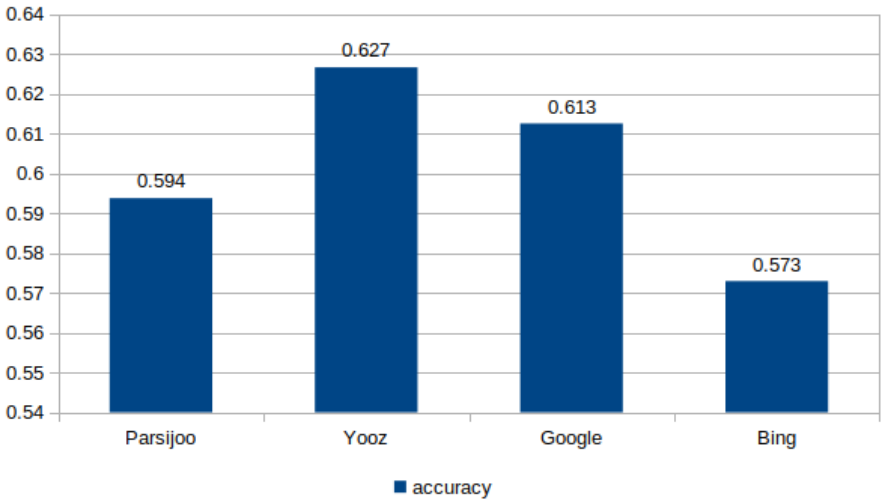


Fig. 3: Tokenizer evaluation

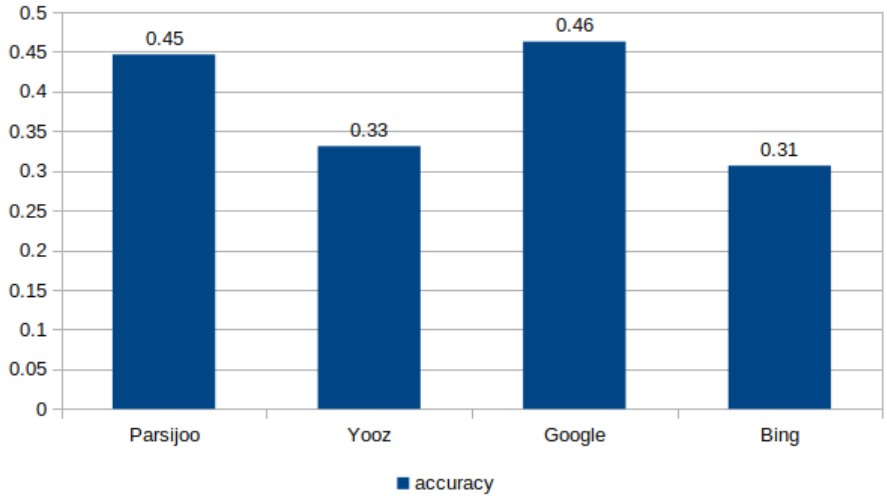
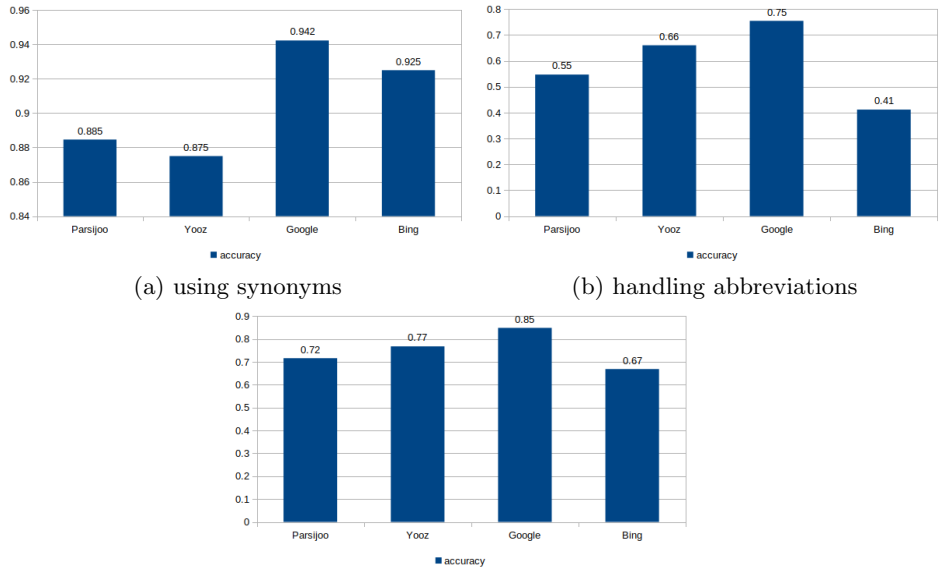


Fig. 4: Spell Correction's evaluation

as it is expected, international search engines achieved better results than the two national ones. In figure 5:a, Google and Bing are meaningfully better than the other two that can be a sign of having a much larger amount of data and undertaking a more sophisticated set of algorithms to expand queries.

Handling abbreviations same as expanding a query using synonyms, the amount of indexed data plays an important role to distinguish abbreviations. Moreover, results show that access to a huge amount of data is crucial but not enough to design a robust system. Figure 5:b, shows that Google and Yooz have a close and acceptable level of supporting abbreviations in a query. Bing's accuracy reveals that regardless of its access to a huge number of indexed pages, it may suffer from not having a sophisticated algorithm for detecting abbreviations.



(a) using synonyms

(b) handling abbreviations

(c) sum up

Fig. 5: Query expansion evaluation

The overall overview given by Figure 10 shows that in general handling synonyms is addressed more than abbreviations in search engines. Bing's Abbreviation's low score vanished the expansion's high score using synonyms. On the other side, two national search engines compensate their score by supporting abbreviations much better than Bing.

Query Analyzer figure 6 shows that Google has the best overall query processor component among the evaluated search engines. It confirms that most of the

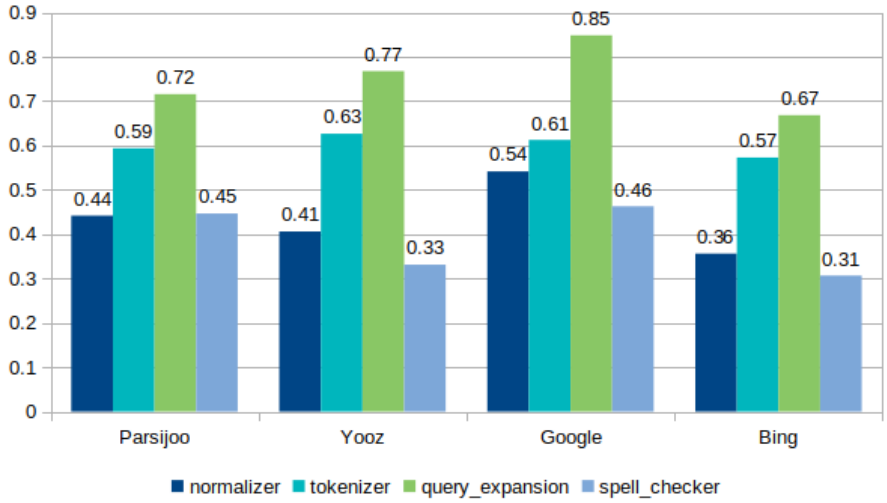


Fig. 6: Phase1 total scores

fundamental search engine's tasks can be addressed using language-independent methods. Additionally, Yooz, Parsijoo are at a similar level at this phase with a mere difference in their total score of some tasks. Bing gained the lowest score in all the tasks, however, its scores are generally close to other search engines.

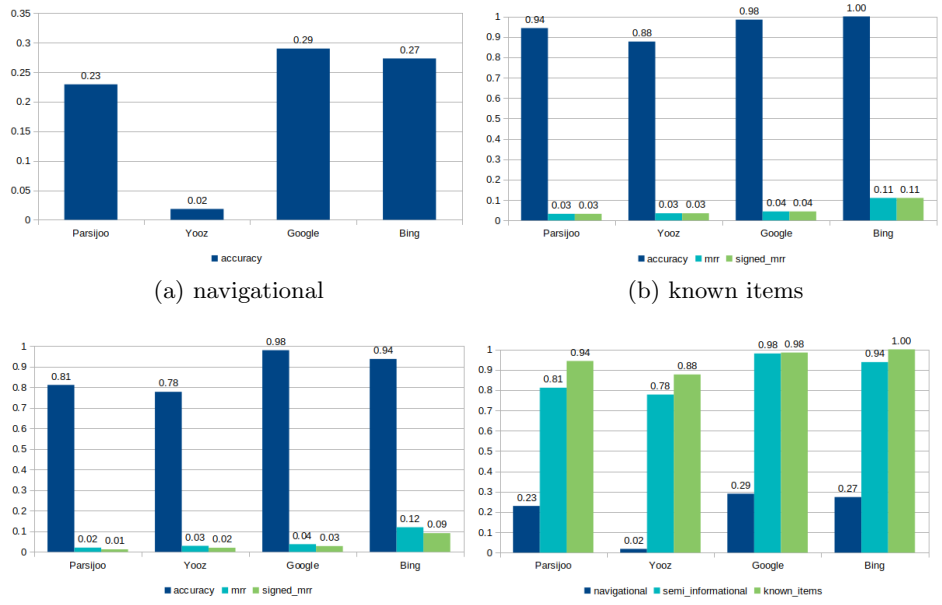
4.5 Rank

In the second phase, about 100 thousand result pages of about 5 thousand queries were fetched and evaluated. The query types that are designed for evaluating this phase are navigational, known items, and semi-informational queries. Navigational is the type of query that the user knows the correct answer's website domain and searches to find that exact one. Known items are similar to navigational, but the query is asking for knowledge not a website domain. Semi-informational queries may have multiple correct answers, e.g. searching for the recipe of food that might have some similar recipes. The decision network is used to evaluate the relevancy of the search engine's result pages for the last two types of queries. As explained previously, different parts of a document are specified in the network and their scores are aggregated.

Navigational queries score of search engines are illustrated in figure 7:a. Google, Bing, and Parsijoo's result in covering this type of queries are similar to each other. However, Google has a giant crawler, Parsijoo has achieved an acceptable score. On the other side, Yooz is far away from the other search engines. It is too disappointing for a search engine; however, it can be the result of a problem in their web crawler or ranking algorithm.

Known items the accuracy score of all the search engines are in an acceptable range, although Yooz is not as well as others. There is a great difference between the value of Mean Reciprocal Rank(MRR) and signed MRR[7] bars with the accuracy score. It shows that relevant answers to this type of query are not among the top-ranked results of search engines. Thus, although search engines like Bing are doing their best in this type of query, they should improve their ranking algorithm to ameliorate their MRR score.

Semi-informational is the main power of Google by which can attract much higher number of users than other search engines like Yooz and Parsijoo. Figure 7:c states that Google and Bing are similar in terms of their accuracy score; however, a higher MRR and signed MRR score for Bing shows its results' higher quality in contrast with Google's. There is a similar relationship between Parsijoo and Yooz in both explained terms. Furthermore, both international engines outperformed two national ones in terms of all determining metrics.



(a) navigational (b) known items

(c) semi-informational (d) sum up

Fig. 7: Rank evaluation

[7] If an irrelevant page is ranked higher than the relevant results of a query the search engine will receive -1 score for that query.

Rank the comparison of the overall score of the rank component of search engines states that Google has done a great job by making a great difference against other systems. National search engines can build a roadmap based on the current results to improve their systems to gain the ability to answer their users' needs.

5 Discussion and Future work

Related work has some weak points like small query set, narrow range of query types, subjective evaluation, high cost of evaluation and re-evaluation, and not evaluating all the components of search engines. The employed method is an automatic and structured-aware solution that addresses all these problems. The significant attributes of this method are suitability, robustness, low cost and re-usability. Results elicit the true weak-points and strengths of each search engine; not only a relative comparison to each other but also a comparison with the true definition of each task that a search engine should address to provide appropriate answers to their users' queries. Moreover, results show that there is a close correlation with experts manual evaluation and Parsisanj's automatic evaluation.
Final results show that although Google achieved the best average score in both evaluation steps, the overall score of the two national search engines were acceptable in most of the assessment steps. In the query analyser step, the two national engines ranked higher than Bing, but in the rank step, Bing beats them. This shows that using some language-specific methods can work in query analysing improvement, but not necessarily in rank step; however, in rank step the large amount of seen web pages might be the key to gain a higher score.

Designing an automatic system for building the query-set is one of the future works that can be investigated to improve this work. The structure of queries can be learned by text-mining methods, after training a robust model, it can extract new queries using an unlabelled corpus. Moreover, some other evaluation metrics can also be used to analyse studying systems in a better level of detail.

Acknowledgment

We are really thankful for Dr. MohammadAli Abam's cooperation and support in managing this project. Truly, he made a great benefit for accomplishing this project. Additionally, This project was supported by Iran Telecommunication Research Center with project id 901952720.

References

1. Mahmoudi, M., Badie, R., Zahedi, M. S., Azimzadeh, M. (2014). Evaluating the retrieval effectiveness of search engines using Persian navigational queries. 7'th International Symposium on Telecommunications (IST'2014), 563–568.

2. Azimzadeh, M., Badie, R., Esnaashari, M. M. (2016). A review on web search engines' automatic evaluation methods and how to select the evaluation method. 2016 Second International Conference on Web Research (ICWR), 78–83.
3. Shoeleh, F., Azimzadeh, M., Mirzaei, A., Farhoodi, M. (2016). Similarity based automatic web search engine evaluation. 2016 8th International Symposium on Telecommunications (IST), 643–648.
4. Nowkarizi, M., Zeinali, M. (2017). The overlap and coverage of 4 local search engines: Parsijoo, Yooz, Parseek and Rismoun. Human Information Interaction, 4(3), 48–59.
5. Sánchez, D., Martínez-Sanahuja, L., Batet, M. (2018). Survey and evaluation of web search engine hit counts as research tools in computational linguistics. Information Systems, 73, 50–60.
6. Zhang, J., Cai, X., Le, T., Fei, W., Ma, F. (2019). A Study on Effective Measurement of Search Results from Search Engines. Journal of Global Information Management (JGIM), 27(1), 196–221.
7. Wu, S., Zhang, Z., Xu, C. (2019). Evaluating the effectiveness of Web search engines on results diversification. Information Research: An International Electronic Journal, 24(1), n1.
8. Rahim, I., Mushtaq, H., Ahmad, S. (2019). Evaluation of Search Engines using Advanced Search: Comparative analysis of Yahoo and Bing. Library Philosophy and Practice, 1–12.
9. Gul, S., Ali, S., Hussain, A. (2020). Retrieval performance of Google, Yahoo and Bing for navigational queries in the field of "life science and biomedicine." Data Technologies and Applications.
10. Tazehkandi, M. Z., Nowkarizi, M. (2020). Evaluating the effectiveness of Google, Parsijoo, Rismoon, and Yooz to retrieve Persian documents. Library Hi Tech.
11. Givi, H., and H. Anvari. "Dastoore zabane Farsi." (2003).
12. Callan, J. P., Croft, W. B., Harding, S. M. (1992). The INQUERY Retrieval System. Database and Expert Systems Applications, 78–83.
13. Metzler, D., Manmatha, R. (2004). An inference network approach to image retrieval. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3115, 42–50.