

Test-retrieval framework: performance profiling and testing web search engine on non factoid queries

Althaf Ali A, R. Mahammad Shafi

Department of Computer Science, Bharathiar University, India

Article Info

Article history:

Received Sep 24, 2018

Revised Nov 17, 2018

Accepted Feb 1, 2019

Keywords:

Big data cloud computing
Query optimization
Recommendation model
Software engineering
Testing web retrieval system

ABSTRACT

Performance profiling and testing is one of the interesting topics in the big data management and Cloud Computing. In testing, we use test cases composed to different type of queries to evaluate the performance aspects of the information retrieval system for large scale information collection. This test scenario performs the evaluation of retrieval accuracy for all kind of ambiguity and non factoid queries with result set as Training data. This stands difficult to evaluate the retrieval method in order to schedule or optimize the Recommendation and prediction technique of the IR method to the Real time queries. The Queries is considered as requirement specification which has to supply to search engine or web information provider applications for information or web page retrieval. In this paper, we propose a novel technique named as "Test Retrieval Framework" a performance profiling and testing of the web search engines on the information retrieved towards non factoid queries. In this technique, we apply expectation maximization algorithm as an iterative method to find maximum likelihood estimate. We discuss on the important aspects in this work based on Recommendation models integrating domain and web usage, Query optimization for navigational and Transactional queries, Query Result records. The Experimental results demonstrates the proposed technique outperforms of state of arts approaches in terms of set based measures like Precision, Recall and F measure and rank based measures like Mean Average Precision and Cumulative Gain.

*Copyright © 2019 Institute of Advanced Engineering and Science.
All rights reserved.*

Corresponding Author:

Althaf Ali A,
Department of Computer Science,
Bharathiar University,
Coimbatore-641046, Tamilnadu, India
Email: althafa579@gmail.com

1. INTRODUCTION

Testing web Information Retrieval System is active research topic in the area of the software engineering [1]. Nowadays web Information Retrieval System used mostly in the cloud computing and big data management as volume of outsourcing data to backup service is increasing day to day [2]. A great deal of the retrieval system has been devoted to resolve these issues over the years. It has reported that approaches based on the recommendation model and prediction model can efficiently represent the information or web pages to the user queries or keywords. Nevertheless determining the performance of the retrieval system is also an essential part of the information retrieval (IR) and to access the ability to meet the user specification in terms of keyword and queries for search result. Perhaps most obviously user depends on the search engine and information retrieval system in order to perform any kind of information accessing [3].

Testing of web information retrieval system or search engine is organized into two kinds as state of art of approaches. They are system based evaluation and user based evaluation. In the system based evaluation, method is quantified on its ability to retrieve and rank the results that relates to query while user

based evaluation depends on user satisfaction [4]. Mostly System based evaluation stands primary important to generate good information retrieval system. However system based evaluation model is more challenging due to expansion of more users of the data. Thus employment of test cases towards of different type of queries and to its result set has become more important to evaluate the performance of the IR with efficiency and robustness in order to eliminate the pitfalls of the state of art approaches. Furthermore test scenario identifies the retrieval accuracy for all kind of ambiguity and non factoid queries with result set as training data [5].

In this paper, we propose a novel technique named as “Test Retrieval Framework“ a performance profiling and testing of the web search engines on the information retrieved towards non factoid queries. In this work, the contribution are applying expectation maximization algorithm as an iterative method to find maximum likelihood estimate for user query. In addition we discuss on the important aspects based on Recommendation models integrating domain and web usage, Query optimization for navigational and Transactional queries, Query Result records for evaluating methods on different data types.

The remainder of the paper is organized as follows: Section 2 discusses the related works in evaluation of IR methods and its impacts against the performing evolving user queries, Section 3 briefly discusses the proposed technique in terms profiling and testing the IR method and Section 4 presents the experimental results on a number of data sets. Section 5 discusses conclusions and future work.

2. RELATED WORK

There exist many techniques to Information Retrieval Evaluation are designed and implemented efficiently. Each of these techniques follows some sort of effectiveness on the evaluation of the System among few performs nearly equivalent to the proposed model which is described as follows

2.1. RETRIEVAL-Web Based IR Analysis System

It is currently available and ready-to-use on-line Web-based IR Analysis System which offers a high level accuracy range on various data input structures related to dissimilarity distances and classification indexes and thus composing a generic IR evaluation. Parallely the interactive performance analysis over a complete ranking list and failure monitoring such as the binary relevance ranking list scatter plot and the dissimilarity matrix is also enabled in this model. The relevancy of the information retrieved to search queries is computed using data classification and clustering technique [6], [7]. The System uses the principle component analysis for transformation to produce uncorrelated and orthogonal principal components. Also it Transfer a set of correlated variables into a new set of uncorrelated variables [8].

2.2. HAMSTER: Search Click Logs for Schema and Taxonomy Matching for Information Retrieval

In this mechanism, unsupervised matching of schema information extracted from a large number of data sources into the schema of a data warehouse is been established. The matching process is the first step of a framework to integrate data feeds from third- party data providers into a structured-search engine's data warehouse for fast retrieval of data. We utilize technique based on the search engine's click logs and Taxonomy. Two schema elements are matched if the distributions of keyword queries that cause click-through on their instances are similar [9].

2. PROPOSED MODEL

In this section, we design “Test- Retrieval a performance profiling and testing of the web search engine on their information retrieval to non Factoid queries. The proposed model incorporates in exploring decision points, action paths and interest area along exploiting Indexing conditions click through and page links using expected maximization algorithm. Detailed description of design is as follows

3.1. Analysis of Web Crawler

It is used to explore the information starting from a number of seed pages, follows outbound links, and so attempts to gather the entire web. The Information crawled based on several conditions. The information gain [8] is calculated for each information extracted from the web database and the best n information are selected based on the time and updates

$$\begin{aligned}
 G(t) = & -\sum_{c_i} p(c_i) \log p(c_i) \\
 & + p(t) \sum_{c_i} p(c_i | t) \log p(c_i | t) \\
 & + p(\sim t) \sum_{c_i} p(c_i | \sim t) \log p(c_i | \sim t)
 \end{aligned}$$

Where

t is the Time, C is the number of categories

Information is selected into the feature space through probability estimation. Information is also represented as data point

i- {d1,d2,d3}

Where d1,d2,d3 is the datapoint of the information in the particular web page or web content.

3.2. Analysis of Web Indexer

The Web indexer will index the crawled information. Index design incorporates concepts in terms of information with key and value pair to optimize the retrieval rate in finding the retrieval information to the search query [10]. Analysis of the indexing depends on the several factors such as:

a. Merging conditions

Initialize Category Set =0

Set of Data type predicate in RDF graph

$I_C \rightarrow \{i\} < \beta$ where β is discriminability

While {category Set \neq 0}

For key \in Candidate set

Discriminability \leftarrow dis(key, I, G)

If (discriminability < β)

Then include the key in category

Else

Coverage \leftarrow con(key, I, G)

FL(key; IC;G)

Return arg maxkey2key set score[key]

b. Lookup construction

The look up condition depends upon link in the rows of the index table containing value and key for information retrieval. The lookup helps for duplicate eliminate through storing only updates of the information in forms of key and value pair [11].

c. Inverted Index

The index generation of the information is represented in terms of RDFS [12] with an example as follows

There are two types of animals, Male and Female.

```
<rdfs: Class rdf: ID="Male">
```

```
<rdfs:subClassOfrdf:resource="#Animal">
```

```
</rdfs:Class>
```

The subClassOf element asserts that its subject - Male - is a subclass of its object -- the resource identified by #Animal.

```
<rdfs:Classrdf:ID="Female">
```

```
<rdfs:subClassOfrdf:resource="#Animal"/>
```

```
<owl:disjointWithrdf: resource="#Male"/>
```

```
</rdfs:Class>
```

3.3. Analysis of Information Retrieval

The analysis of the information retrieval is computed for testing on following methods for different kind of queries

3.3.1 Analysis of Ranking Algorithm

It will be decide the Result list for the user queries by referring the indexer. Ranking is usually ranked based on the date of publication or based similarity measures between the content [13]. Ranking Algorithm contains huge ranking constraints. Most popular Ranking Algorithm apply term frequency and inverse document frequency mechanism (TF-IDF), term frequency means that documents containing more occurrences of terms also found in the query are ranked higher, and inverted document frequency implies that documents containing query terms that are rare throughout the index are also considered more relevant. The Figure 1 describes the architecture of the proposed model. Another ranking Approach is ranking approach is the Vector Space represents every document as a vector, with every term occurring in the index providing a dimension of the vector space, and the number of occurrences of the term in a information providing the extension in this dimension [14].

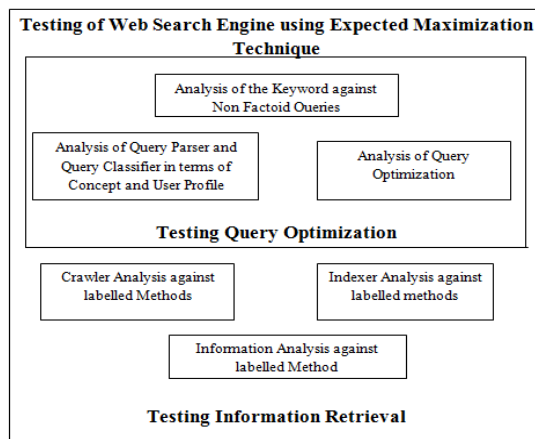


Figure 1. Architecture diagram of the proposed test retrieval framework against search engine optimization

For the query another vector is constructed in the same way; the ranking is then determined by a similarity measure (e.g. the cosine) between the query vector and single document vectors which contains different query types such as Information queries [15] and Navigational queries [16].

3.3.2 Analysis of Link Based Algorithm-Hits and Page Rank

The HITS algorithm retrieves all documents relevant to the query and assigns hub values to outgoing links and authority values to values to those with many incoming links from the link network the most selective and the most selected documents [17].

Page Rank algorithm ranks the information and gives the value for each page based on every link from one page to another page. The ranking is always based on a large number of criteria. The expected maximization algorithm computes the similarity between the search retrieval results against page relevancy and time of retrieval [18].

3.3.3 Analysis of User Log Data-User Characterization

Analysis of user log is essential to make discrimination between the various users. The web information retrieved for various kinds of user queries is stored in cache and session is utilized. Classification of server log of the user is applied through supervised and unsupervised technique through characteristics. Analysis is applied in terms of log parser, after log parsing, it is computed with patterns of the user profiles.

3.4. Query Classification

The queries are traditionally divided into navigational (user looks for a specific web page known or supposed to exist), transactional (not surprisingly, in this case the user looks to perform a transaction like buying or downloading) and informational, which should need no further explanation. They are classified based on the distribution. The keyword analysis is carried out for keyword optimization on non factoid queries, where query parsing and query optimization is also processed towards achieving conceptual and semantic queries to search engine to produce gain on information retrieval. Evaluation has to be focused on Information queries instead of other queries [19]. Once the user submits his query, he is presented with the result page with set of results in order. Each result consists of three parts: a page title taken directly from the page's <title>tag, a so-called "snippet" which is a query-dependant extract from the page, and a URL pointing to the page itself.

3.5. Expected Maximization

It is Iterative method for learning probabilistic categorization model for web search engine evaluation against various kinds of queries. Initially assume random assignment of examples to categories. Learn an initial probabilistic model by estimating model parameters θ from this randomly labeled retrieved result for queries provided [20]

- Expectation (E-step): Compute $P(c_i | E)$ for each example given the current model, and probabilistically re-label the examples based on these posterior probability estimates.
- Maximization (M-step): Re-estimate the model parameters, θ , from the probabilistically re-labeled data.

Let's considered a completely labeled Query Q and Result Set R for queries and Q_m query processing methods and Information retrieval methods I_R , and randomly select a subset as D_K . Also use the set of unlabeled information in the EM procedure. Correct classification of a information to the query is:

Concealed class label = class with largest probability

Accuracy with unlabeled information set or result set > accuracy without unlabeled information retrieved. Criteria for initial Iteration is given by:

$$\Pr(c_d | d) = 1 - \varepsilon \text{ and } \Pr(c' | d) = \varepsilon / (n - 1) \text{ for all } c' \neq c_d$$

Let the class probabilities of the labeled and indexed information is taken to re iteration based on the features extracted and features grouping for evolutionary features for computation with labeled method for search engine evaluation for all kinds of queries. Keeping labeled set of same size Laplacian law for regularized class is given by

$$\theta_{c,t} = \frac{1 + \sum_{d \in D_c^k} n(d, t)}{|W| + \sum_{d \in D_c^k, \tau \in d} n(d, \tau)}$$

Re iterate $\Pr(c|d)$, for each feature for each query and outputs the result set. It estimates class-conditional distribution which includes information from D. Once a new model is trained, it replaces one of the existing models in the expected maximization. The candidate for replacement is chosen by evaluating each model on the latest training data, and selecting the model with the worst prediction error. This ensures that we have exactly L models in the ensemble at any given point of time. In this way, the infinite length problem is addressed because a constant amount of memory is required to store the maximum equivalent data. The concept-drift problem is addressed by keeping the expected information up-to-date with the most recent concept.

Algorithm 1: Test Retrieval Analysis

Input: Information Retrieval Methods, Query processing methods and Different Categories of Query, Labelled information with queries

Output: Evaluation of the web search Engine performance

Process:

1. Initialize Instance pair IP
2. Where $C_p \rightarrow 0$
3. Resultant Query Set
4. For each Instance Pair $IP_1 \in IP$
5. Generate query Q
6. Where $Q = \{LT_1, i\}$
7. $Q = \arg \text{Max} (Q) \in IP$
8. Return Q
9. Categories = {C1, C2, C3}
10. Instance of Source 1= {I₁1, I₁2,....}
11. Instance of Source 2= {I₂1, I₂2,....}
12. Matching instance is carried out using Expected Maximization
13. Where r = Support value 1 or 0

$$r = y \frac{w^T x + b}{\|w\|} \tag{1}$$

14. 0 Represent Non matching

15. 1 Represent Matching

3. EXPERIMENTAL ANALYSIS

In section, we describe the experimental results of the proposed framework against the existing approaches. The analysis of the search engine can carried out with 2 real datasets YAGO and DBPedia.

YAGO is a Semantic knowledge base, in which entities, facts, and events are anchored in both time and space. YAGO2 is built automatically from Wikipedia, GeoNames, and Wordnet. It contains 447 million facts about 9.8 million entities. Human evaluation confirmed an accuracy of 95% of the facts in YAGO [21].

DBpedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to ask sophisticated queries against datasets derived from Wikipedia and to link other datasets on the Web to Wikipedia data. We describe the extraction of the DBpedia datasets, and how the resulting information is published on the Web for human-and machine-consumption [22].

4.1. Evaluation of Test Retrieval Framework on Set Based Measures

4.1.1 Precision

The precision of a retrieval system for a certain query is the proportion of results that are relevant.

$$\text{Precision} = P = \frac{\text{RelevantretrievedResult}}{\text{OverallRetrievedResult}}$$

The Figure 2 describes the search engine evaluation against precision on existing and proposed technique, where proposed value proves the test application predicts the fault and its normal functioning of the search engine through relevant data retrieved. It is a scalar metric used as measure the performance of the system over all other relevant results through ranking.

4.1.2 Recall

The Recall of a retrieval system for a certain query is the proportion of relevant results that have been retrieved.

$$\text{Recall} = R = \frac{\text{RelevantretrievedResult}}{\text{RelevantResultindatabase}}$$

The recall and precision measures are generally inversely proportional: if a retrieval system returns more results, the recall can only increase (as the number of relevant results in the database does not change), but precision is will be decreased.

The Figure 3 describes the search engine evaluation against Recall on existing and proposed technique; in this case proposed system generates the high possible results compared with other possibilities in information retrieved to different class or categories of queries.

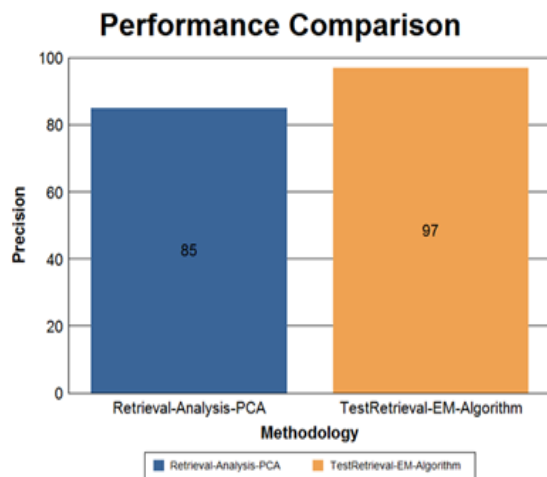


Figure 2. Performance analysis of the search engine evaluation against the precision

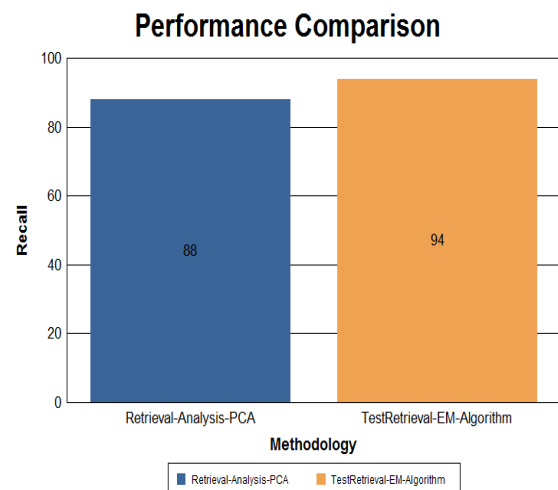


Figure 3. Performance analysis of the search engine evaluation against the recall

4.1.3 F Measure

It is a measure of a test's accuracy and is defined as the weighted harmonic mean β of the precision and recall of the test. F measure is given by:

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

The f measure performance outcomes is described in the Figure 4 towards employing the proposed and exiting technique, among both proposed system yields better results as described in Figure 4

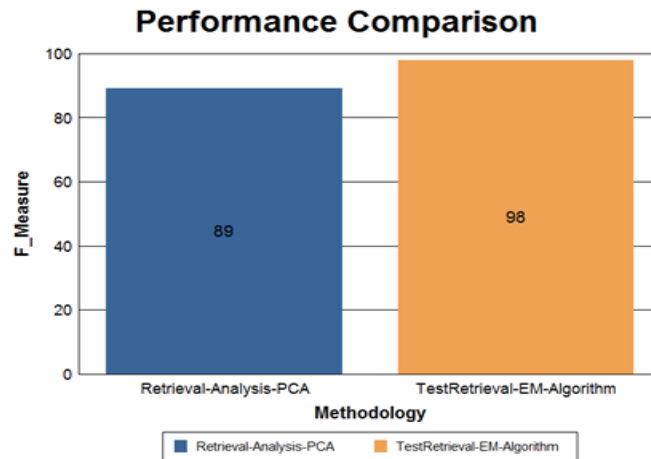


Figure 4. Performance analysis of the search engine evaluation through f measure

4.2. Evaluation of Test Retrieval Framework on Rank Based Measures

4.2.1 Mean Average Precision

It averages mean precision of the multiple queries. MAP considers the precision at every relevant result in the result list for queries. The precision is averaged by dividing the sum of precisions by the total number of relevant results.

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} + \sum_{k=1}^{m_j} precision(R_{jk})$$

The MAP outcomes described in the Figure 5 which shows that proposed mechanism yield better results compared with existing system

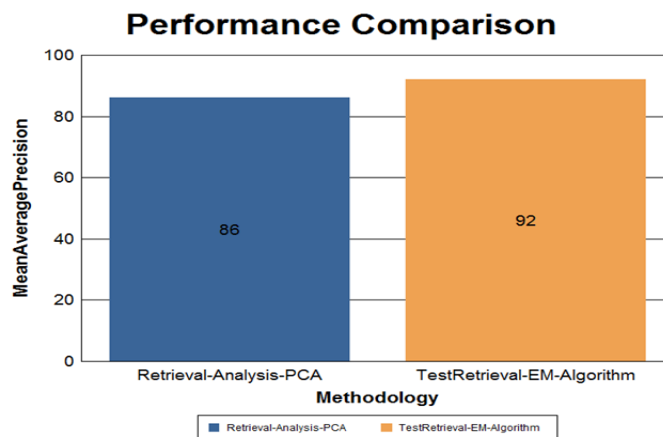


Figure 5. Performance analysis of the search engine evaluation against mean average precision

4.2.2 Cumulative Gain

It is measures the graded relevance to detect the usefulness or gain from examining the information of the retrieval system. Cumulative Gain is given by:

$$CG = 1/\log(\text{Rank})$$

Where, Rank determines usefulness of the information for the queries provided.

The Cumulative gain is described in the Figure 6 states that ranking of the relevancy information on the proposed model. It explains the effectiveness of the proposed function on determining the faults.

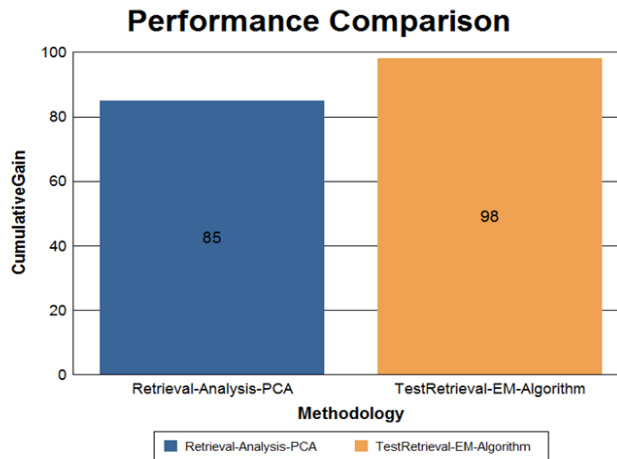


Figure 6. Performance analysis of search engine evaluation against the cumulative gain

The Table 1 describes the performance analysis of the search engine evaluation on the technique described in this research against various measures to compute its strength. The proposed algorithm discovers the web search performance and fault condition against various measures [23].

On experimental evaluation of proposed framework, it has proven that it can used to evaluate leading search engines like GOOGLE, Yahoo and Bing. It provides accurate result in all terms of evaluating parameters as compared with state of art approaches.

Table 1. Performance Analysis of the Search Engine Evaluation

| Technique | Precision | Recall | F measure | Mean Average Precision | Cumulative Gain |
|---|-----------|--------|-----------|------------------------|-----------------|
| Retrieval Analysis – PCA (Existing) | 85 | 88 | 89 | 86 | 85 |
| Test Retrieval – EM Algorithm (Proposed) | 97 | 94 | 98 | 92 | 98 |

4. CONCLUSION

We designed and implemented Test Retrieval Framework which is a performance profiling and testing of the web search engines on the information retrieved against non factoid queries. The testing architecture analysis each component of the query and information retrieval mechanism using expected maximization process. The EM algorithm is the iterative algorithm which determines the maximum likelihood of the analysis of search engine against labelled information. It is the first work to evaluate the search engine in terms of query model, crawler model, Indexing model, ranking model and Information retrieval model in terms of concept based mining. The performance analysis of the proposed model is computed against different measure. In future work, incorporation of several web mining mechanism will optimize the performance of information retrieval.

REFERENCES

- [1] Raghavan, V.; Bollman, P.; & Jung, G.S. A critical investigation of recall and precision as measures of retrieval system performance. *Communication of the ACM*, 1989, 7(3), 205-29.
- [2] Omaioulo, Nicholas G. & Packer, Joan G" An analysis of Internet search engines: Assessment of over 200 search queries, The list of queries used is in: Quantitative Analysis of Five WWW 'Search Engines'. *Computers in Libraries*, 1996 16(6), 58 to 62.
- [3] Lempel, R.; & Moran, S. Predictive caching and prefetching of query result in search engines. In *Proceedings of the 12th World Wide Web Conference (WWW2003), Budapest, Hungary*, 2003.
- [4] Hamid Sadeghi. "Automatic Performance Evaluation of Web search Engines using judgements of Meta search Engines", Online Information Review, ISSN: 1468-4527, *Emerald Publishing Limited.*, 2011, 12(6)957-971.
- [5] D. Kelly, "Methods for evaluating interactive information retrieval systems with users," *Found. Trends Information Retrieval*, 2009, 3(1), 1–224 [Online]. Available: <http://dx.doi.org/10.1561/1500000012>
- [6] M. Agosti, E. Di Buccio, N. Ferro, I. Masiero, S. Peruzzo, and G. Silvello, "DIRECTIONS: Design and Specification of an IR Evaluation Infrastructure" *Berlin, Heidelberg: Springer Berlin Heidelberg*, 2012, 88–99. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33247-0_11
- [7] Lewandowski, D., Drechsler, J., & Mach, S. Von. "Deriving Query Intents from Web Search Engine Queries. *Journal of the American Society for Information Science and Technology*, 2012 63(9), 1773–1788,
- [8] Lewandowski, D. & Höchstätter, N. "Web Searching: A Quality Measurement Perspective. In A. Spink & M. Zimmer (Eds.), *Web Search: Multidisciplinary Perspectives*, Berlin, Heidelberg: Springer, 2008, 309–340.
- [9] Lewandowski, D., & Spree, U. "Ranking of Wikipedia articles in search engines revisited: Fair ranking for reasonable quality" *Journal of the American Society for Information Science and Technology*, 2012 62(1), 117–132.
- [10] J. Urbano, "nfire a framework for information retrieval evaluation in. net, <https://github.com/julianurbano/nFire>, accessed: 2016-02-18.
- [11] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel, "Evaluatir: An online tool for evaluating and comparing ir systems," in *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR'09. New York, NY, USA: ACM*, 2009, 833–833. [Online]. Available: <http://doi.acm.org/10.1145/1571941.1572153>
- [12] X. Benavent, A. Garcia-Serrano, R. Granados, J. Benavent, and E. de Ves, "Multimedia information retrieval based on late semantic fusion approaches: Experiments on a wikipedia image collection," *IEEE Transactions on Multimedia*, 2013, 15(8), 2009–2021.
- [13] J. Kekäläinen, "Binary and graded relevance in ir evaluations comparison of the effect on ranking of ir systems," *Inf. Process. Manage.*, 2005, 41(5)1019–1033 2005. [Online]. Available: <http://dx.doi.org/10.1016/j.ipm.2005.01.004>
- [14] D. Lewandowski, "The retrieval effectiveness of search engines on navigational queries," *Aslib Proceedings*, vol. 63, no. 4, pp. 354–363, 2011.
- [15] S. Qiao, T. Li, H. Li, Y. Zhu, J. Peng, and J. Qiu, "SimRank: A page rank approach based on similarity measure," *Proceedings of 2010 IEEE International Conference on Intelligent Systems and Knowledge Engineering, ISKE 2010*, 09, 390–395
- [16] T. Sakai, "On the reliability of information retrieval metrics based on graded relevance," *Information Processing and Management*, 2007, 43(2), 531–548.
- [17] W. Magdy and G. J. F. Jones, "Examining the robustness of evaluation metrics for patent retrieval with incomplete relevance judgements," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010 63, 82–93.
- [18] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, "YAGO2: A spatially and temporally enhanced knowledge base from wikipedia," *Artif. Intell.*, 2013
- [19] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives, "DBpedia: A nucleus for a web of open data," in *ISWC/ASWC*, 2007.
- [20] A. Nandi and P. A. Bernstein, "HAMSTER: using search clicklogs for schema and taxonomy matching," *PVLDB*, 2009.
- [21] Andre Andre " Medical Information System Services and Patient Queue Management to Enhance Services Quality in Indonesia" *International Journal of Electrical and Computer Engineering*, Vol:9, Issue:3, 2019.
- [22] Soukaina Cherif Bourki Semaili, Mohammed Essaid Riffi, Fayçal Chebihi "Parallel Hybrid Chicken swarm optimization for solving the Quadratic assignment problem" *International Journal of Electrical and Computer Engineering*, Vol:9, Issue:3, 2019.
- [23] Maslina Abdul Aziz, Jemal H. Abawajy, Morshed Chowdhury " Scheduling Workflow Applications with Makespan and Reliability Constraints" *International Journal of Electrical and Computer Engineering*, Vol:9, Issue:3, 2019.