Achieving High Recall and Precision with HTLM Documents: An Innovation Approach in Information Retrieval

Ammar Al-Dallal[1] and Rasha S. Abdulwahab[2]

Abstract—**Information retrieval techniques become a challenge to researchers due to huge growth of digital and electronic information. Researchers are attending this area by developing different techniques to enhance precision and recall of retrieved documents. This paper presents an information retrieval system that has promising results in terms of recall and precision. These results are achieved via developing an improved inverted index for the document set and by developing an enhanced evaluation function to evaluate the retrieved documents in response to user query. Results are compared with two well known techniques applied in IR domain which are Okapi-BM25 and Bayesian interface network model and show that precision and recall of the retrieved documents by the proposed method outperforms these two techniques.**

***Index Terms*— evaluation function, term distance, terms proximity.**

I. Introduction

Paying particular attention to the importance of providing suitable information for the user's needs, many studies highlighted the importance of measuring the effectiveness of the retrieval results. Toward this accessing to the relevant information which is provided by World Wide Web has been examined by many pioneers in this field [4]. Retrieving relevant information is a complex process and the complexity is further increased by the fact that more and more of this information appears in natural language and not in structured formats [10].

Information Retrieval (IR) is primarily devoted to extracting relevant information rather than data. The study of IR techniques has increased since the advent of the World Wide Web, but still web users suffer from two problems when trying to retrieve useful information. One of them is that many of the highly ranked retrieved documents are not related to the user query. On the other hand, there still many related documents which are not retrieved [16]. For this reason, many paradigms and models have been developed to solve the IR problem. .

The paper is organized as follows: In section 2, we discussed IR problem while the objectives are illustrated in section 3. Section 4 discusses related work. Document representation is introduced in section 5. In section 6, implementation of proposed IR system is introduced. The results are analyzed in section 7. Conclusions of the results and directions for future work are given in last section.

[1] School of Information Systems Computing and Mathematics, Brunel University, U.K, Email: Ammar.AlDallal@brunel.ac.uk.
[2] College of Information Technology, Ahlia University, Bahrain, Email: rasha_sh_abdul-wahab@ahliauniversity.edu.bh.

II. Problem Statement

Given a user query, there is a need to have an IR system that is able to retrieve *all* and *only* relevant documents to the user query. The performance of the developed IR is evaluated using two well known measures which are precision and recall.

III. Research Objective

The objective of this paper is to explain how certain indexing schema can affect the efficiency of the proposed IR system in addition to develop new evaluation function that is able to filter the relevant documents based on user query and retrieve them at high rank.

IV. Related Work

Modern technologies including information and digital technologies have given rise to develop an information retrieval system to acquire the desired information in response to the user queries. Building such systems will help users to efficiently acquire desired information. Of course, it is well known that developing such system is not effortless. For this reason, many challenges have been shaded some light on the current system to increase the effectiveness of IR system.

In information retrieval system (IR), the query is issued and a set of documents that are deemed relevant to the user queries are retrieved. Consequently, the retrieved documents are ranked by applying some empirical evaluation measures. For these reasons, many researchers have suggested several measures that have been used to rank the retrieved documents.

Salton and Buckley [8] are one of the leading authorities in the area of information retrieval. They described in their work some statistical measures that are used to evaluate how important a word is to a document in a collection. *tf-idf* (term frequency inverse document frequency) was proposed and examined in their research. In additional to these factors, authors use frequency of most frequent term in the collection. A number of term-weighting experiments are described and are tested with six document collections of varying size, covering different subject areas.

Another study was proposed by Kim and Zhang [14] to learn several factors which are used to rank the retrieved documents. Genetic algorithm (GA) of HTML was the suggested structure in their work. GA is applied in their work

ISBN: 978-988-19251-5-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

to adopt multiple factors of HTML tags to re-rank the documents retrieved by standard weighting schemes. The proposed method has been tested on artificial text sets and a large-scale TREC document collection. The experimental results of their work show an improvement in average precision when using tagged information over non-tagged information.

In Persin et al [21] have been proposed an evaluation technique that uses early recognition of which documents are likely to be highly ranked in order to reduce costs(i.e., cpu time, memory usage). The main objective of this technique is to speed up document retrieval by avoiding processing all indexes that are relevant to the given query. They observed, ordering the inverted lists by decreasing within-document frequency is effective by keeping the first part of each list with high frequencies, and ignoring the rest. Consequently, an appropriate evaluation strategy on the resulting lists has been adopted such that the documents with the highest scores will be identified without scanning the entire lists. Wall Street Journal articles (extracted from the TREC data) have been used to evaluate their technique. The experimental results show the proposed techniques have been maintained retrieval effectiveness with reducing the sources used.

One of the well known evaluation function used in IR is Okapi-BM25. This function is used in many researches and found to be consistently performing very well in TREC competitions [14]. This formula is defined as follows:

$$f(d) = \sum_{T \in Q} \frac{(k1+1)\times tf}{k1 \times ((1-b)+b\frac{length}{length_{avg}})+tf} \times log \frac{N-df+0.5}{df+0.5} \quad (1)$$

where tf is the term frequency, df is the number of documents in the collection referencing the considered term while length and lengthavg are the document length and the average document length in the collection respectively. k1 and b are constants set to 2 and 0.75 in the study done in [14].

By looking at the techniques mentioned above and the evaluation functions used, it is noticed that most of them are using vector space model to present the document collection and are applied on non-structured documents. Moreover, these approaches are focusing on limited factors in term weight or document evaluation formulas. At this point, the contribution of our work to IR is to add several effective factors to evaluate the document and to adopt HTML documents that are indexed by using new inverted index schema.

V. Document Representation

In the IR system, the adoption of effective way to represent documents has greatly influenced the scientists' thought. Actually, the documents that will be evaluated by IR system can be either plain text, semi-structured (i.e., HTML (HyperText Markup Language) documents) or structured. Because most of web-documents are written in HTML [12], this format was adopted for implementing of our proposed system.

The primary concern in representation is how to select proper index terms [11] and which indexing model to be implemented. Many indexing models were developed for this purpose such as Boolean indexing model [7], vector spacing model [2], [5] and [6] and latent Symantec indexing model [17]. However, several limitations are raised with these models which are:

- These models require large space to store the index,
- Require long time to retrieve the needed terms, and
- Documents also have limited information to store.

To overcome these drawbacks, a well known indexing schema is developed which was chosen for implementation of our system. This indexing schema is called inverted index [3]. It is perhaps the most important index method used in search engines as stated by Liu [3].

VI. The Proposed System Framework

In the proposed system which we call it Information System With Innovative Evaluation Function (IRWIEF), all documents in the search space will be evaluated by using the proposed new evaluation function. Consequently, the retrieval results are ranked in descending order. The displayed results are those documents which have at least half of the keywords queried by the user. That's because the documents having less than half of the keywords queried by the user are most likely not related to the entered query.

The developed system consists of four main units outlined as shown in Figure 1. Each unit in the proposed system will be used to retrieve the relevant information in response to the user queries. These units are explained in the following.

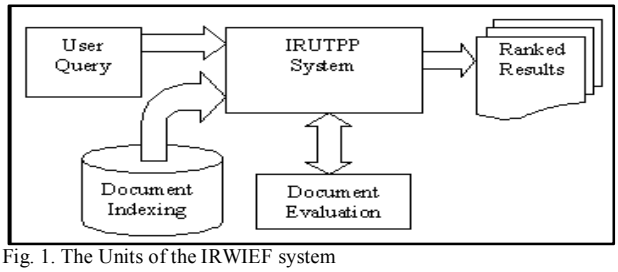

Fig. 1. The Units of the IRWIEF system

A. Document Indexing

One of the important reasons for indexing documents is the difficulties involved in implementing document-term weighting [20]. This indexing schema not only allows different retrieval of documents that contain query terms, but also very fast to build. In its simplest form as described in the algorithm 1, the inverted index of a document collection is basically a structure that attaches each distinctive term with a list of all documents that contain the term.

The power of the proposed system is coming from the used indexing technique. Indexing of a set of document is created by enhancing the known inverted index. In the traditional inverted index [19], the schema is created by appending the position of word in the collection and the document identifier. Word position data is a list of offsets at which the words occur in the document. Such occurrence information (i.e. document ID and word position data) for each word is expressed as a list, called "inverted list". This schema is enhanced by

ISBN: 978-988-19251-5-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

encapsulating extra information related to the indexed term. This information includes term weight, term frequency. The term weight is estimated to represent the number of times the term appear within the document and the position or offset of the term within the document within each sentence. This offset will be used in calculating the term distance.

Algorithm 1 Enhanced inverted index

```
1: while there are more documents need to be process do
2:   while there is a word in the document do
3:         Read the document word at a time until the whole document is
           read. Split the string into tokens.
4:         Remove stop words
5:         Obtain HTML tag weight of the word.
6:         If there is no record for this word for this document
7:                 insert new record including word, position offset,
                   weight and document reference.
8:         Else
9:                 Update existing record of word and document by
                   modifying the new word weight, concatenate the new
                   position.
10:        if there is a record for same document
           then
11:                increment document size in that record
12:        else
13:                insert new record for this document and set document
                   size to 1.
14:        end if
16:    end while
17: end while
```

While creating the index, the offset or position of each term within the document is stored in the index. This position is used to determine the two factors related to term proximity. These factors are the minimum term distance (MTD) between all query keywords within the document and the first appearance (FA) of the minimum distance obtained by MTD.

An Oracle database is used for storing and accessing the enhanced inverted index with the aim of facilitating document retrieval in the face of relevant queries. The data in our proposal are organized with two tables. Each distinct term is stored in a table together with their offset, weight, frequency per document and pointers to the document's identifier. While, another table will be used to store the details of each document includes document size, frequency of maximum frequent term, document weight and number of unique terms. These values for the two tables are computed only once and are independent of a user query. As we have described in [1], the usual approach of building the inverted index was developed by using the link list data structure and was implemented by C++. With this technique, the IR system was failed to process 2000 documents in addition to increase the time for processing these documents. Actually, it is particularly important in IR system processing large volumes of data which will be achieved in the technique used in this paper.

B. Document Evaluation

For our ranking technique, the decision about whether to take or reject a document depends only on the computed value by the proposed evaluation function. In order to understand the new evaluation function along with its components, local and global factor concept together with the concept of term weight will be described.

Local Factors Verses Global Factors Index

The core of this system is the adoption of new evaluation function. This function is a performance measure or reward function that measures the relativity of documents to the user query. To define the proposed evaluation function for information retrieval, we need to define (1) local factors; and (2) global factors. Local factors are those obtain from the document under consideration such as document size, number of unique terms within the document and total number of a specific term within the document. On the other hand, global factors are those obtained from the search space such as total number of documents in the space, total indexed terms and total number of a specific term within the search space [12].

Practically, when the number of documents in the collection is unlimited like the web, local factors are used. That is because; documents' evaluation is done independently of others in the same set. On the other hand, global factors are used when the documents in the collection is limited. In this case the retrieved documents are relatively relevant in comparison to the documents within the set and may not be purely relevant. However, using global factors requires processing of all documents to extract the factors that need to be included in the evaluation function. This requires extra load on the system and additional time for processing. For these reasons, only local factors are included in proposed evaluation function.

Term weight

Term weight is the second concept that needs to be clarify in this section. In HTML documents, tags play an important role in emphasizing the importance of some terms within a document such as:

- Location of the term within the document such as terms appearing in the title.
- Header and anchor text - the text segments which serve as hyperlinks to other documents [14].
- Terms that have specific format such as bold, italic and underlined.

Each HTML tag has a specific weight depending on its importance [1] . The terms will have weights ranging from 1 if it is normal text within body tag to 6 if it is in the title tag. If the term appears in nested tags then its weight is summation of these tag weights. This weight will be one of the factors used in the new evaluation function.

The Proposed Evaluation function

One of the main measures in the IR system is the evaluation function which is used to evaluate the relevance of documents. Indeed, two new concepts are used in the proposed evaluation function along with HTML term weight. In particular, keyword proximity (i.e., term distance) and number of unique query keywords that exist in the document represent the proposed concepts in our evaluation function. Keyword proximity is computed by determining minimum term distance (MTD) between all query keywords within the

ISBN: 978-988-19251-5-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

document and first appearance (FA) of the minimum distance found from previous term.

***Definition** 1:* Minimum Distance (MTD) Initially MTD is a large value greater than the document size. Otherwise, a shorter distance value is assigned if the possibility of distance between query keywords is found. Indeed, keeping the value of MTD large if there is some keywords are missing from this document.

***Definition** 2:* First Appearance (FA) of the minimum distance is computed depend on the value of MTD. This term is computed by dividing the document size over the average distance of MTD. Consequently, the resulting value is divided again by the document size.

The reason behind using MTD is that related terms are most likely appear close to each other, hence, if the MTD is found to be small, that means the document is most likely to be related to user query and vice versa [20]. While FA indicates the relativity of document to the query if the query keywords appear close to the beginning of the document such as in the title, header or in the first paragraph. Higher values for this factor indicate higher relativity.

By using the above notations (factors), the proposed evaluation function is represented as in 2. Table I illustrates the terminologies of equation 2.

$$f(P) = a\frac{\sum_{i=1}^{K} k_{iu}}{K} + b\frac{\sum_{i=1}^{K} k_{iu} - 1}{\sum_{i=1}^{K-1} \min(d_{i,i+1})} + c\frac{F}{\frac{avg(\sum_{i=1}^{K-1} \min(d_{i,i+1}))}{F}} + d\log\left(\frac{\sum_{i=1}^{K} w_i}{\sum_{i=1}^{K} k_i}\right) \quad (2)$$

TABLE I
DESCRIPTION OF TERMINALS OF FORMULA 2

Terminology	Description
k_{iu}	Unique query keyword exists in the document
K	Total number of unique keywords in the query
$d_{i,i+1}$	Distance between term i and term i+1of the query terms
F	Document size (total number of terms in the document)
w_i	Weight of term i in the document as per table 1
a, b, c and d	Weighting factors for each component

This function has four components. First one is the unique number of the keywords of the query that exist in the document. In another words it reflects how many of the query terms are exist in the document. This component has maximum value of one in case of all query keywords exist in the document.

Second component (i.e., MTD) employs the minimum distance between query keywords in the document. It is evaluated by subtracting the number of unique keywords in the document by one. Consequently, the resulting value is divided by the shortest distance between query terms in the document. The reason of subtracting one in the dominator is that the minimum distance between j terms is j-1. In best situation this component returns one if the document contains all query keywords and they occur in consecutive positions at least once within the document.

Example 1: to understand the way of calculating component 1 and 2, assume we have document having the following string of words of length 10:

ABCDEFGHCIJD

and assume the query value is CDF, while the offset of these words within the document are 3, 4, and 6 respectively.

The MTD and the average position of minimum distance between query terms are calculated as follows:

MTD= min(C, D) + min (D, F) = 1+2 = 3

$$avg\left(\sum_{i=1}^{K-1} \min(d_{i,i+1})\right) = 12/\,[(3+4+6)/3]\,/12 = 0.231$$

Third component depends on the position of first occurrence of MTD. It is evaluated by dividing the document size F by the average position of minimum distance between query terms (i.e.: $avg(\sum_{i=1}^{K-1} \min(d_{i,i+1}))$). Consequently, the resulting value is divided again by the document size. In the case where the keywords are very close to the beginning of the document, the value of this component will be high. Also, this factor will return a value close to one, if all query keywords are exist and the positions of these keywords are close to the beginning of the document.

$\log\left(\frac{\sum_{i=1}^{K} w_i}{\sum_{i=1}^{K} k_i}\right)$ represents the fourth component in the proposed evaluation function which is the average weight of query terms in the document. This component reflects the importance of the query terms within the document in terms of HTML. The maximum value of this component is one that is achieved if the summation of terms' weight is 10 times greater than the frequency of these terms. Actually, the log function is used to control the upper limit of this component.

According to the analyzed results, document is considered to be relevant and included in the solution if it has a minimum value of 0.6. Best documents returned by this formula are the ones that have the following:

- All terms of the query keywords,
- In the same sequence of the query,
- In consecutive positions within the document, and
- Appearing at the beginning of the document and within an HTML.

VII. Experiments and Results

This section describes the experimental setup used. Actually, the proposed system is examined on a document set of 8350 using 15 different queries; the length of the queries ranges from 2 to 5.

A. Data Set Description

The data set used in this system is Carnegie Mellon University data set which is HTML documents collected from computer science departments of various universities in January 1997 by the World Wide Knowledge Base project of the CMU text learning group consisting of 8282 documents [18]. This set is grouped into seven categories, named: course, department, faculty, project, staff, student and others in additional to 60 web documents downloaded arbitrary from the Web.

ISBN: 978-988-19251-5-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

B. *Coefficients Setting*

In most of our experiments, we gave high importance to the components that reflect the term distance and number of referenced keywords of the query. Therefore, multiplying each component with a reasonable weighting coefficient has been examined in this paper. Actually, 0.3 is set for coefficients a, b and c to give high importance to the components that use term distance while 0.1 is set to d so that lower importance is given to the weight component.

C. *System's Evaluation*

The results of the proposed system are evaluated by using precision and recall measures. Recall is defined as the percentage of relevant retrieved documents to the total number of relevant documents. While precision is defined as the percentage of relevant retrieved documents to the total number of retrieved documents. A combined method of recall and precision is used to evaluate the validity of the proposed System. Precision versus recall is also used to evaluate the validity and the efficiency of the proposed system. This paper used two different methods. The first method [15] is Precision at Rank N (P@RankN) and Recall and Rank N (R@RankN) where N is multiple of 10. In this method, the retrieved documents are ranked in descending order based on evaluation value and the average of precision and recall are calculated. While in [4] the precision is obtained when recall is multiple of 10%.

D. *System's Performance*

An experimental study was conducted to examine and evaluate performance of the newly evaluation function. Two known evaluation functions using 15 different queries are used here. One of them is OKAPIBM25 [13] which consistently performing very well in TREC competitions [16]. While, the second evaluation function is Bayesian interface network model [14]. The results from these two functions are compared with the proposed evaluation function.

E. *Hypothesis Statement*

To examine the validity of the proposed system, the performance of the proposed system is investigated against both the OKAPI- BM25 and Bayesian interface network model. The effect of proposed indexing schema in the proposed IR system is also observed. The following hypotheses will be examined in this study:

Hypthesis1: The newly evaluation function achieves better performance than OKAPI- BM25 and Bayesian interface network model.

Hypthesis2: The newly developed indexing schema achieves better performance than the indexing schema proposed in [1].

F. *Results Analysis*

Hypothesis 1 is analyzed and examined in this section. Figure 2 shows the average precision (78%) for the proposed evaluation function is larger than of the other methods for the first 10 ranked documents. Moreover, small value of average precision for both models (45%) was found for the first 10 retrieved documents. The improvement in precision (66%) of the proposed evaluation function than that of the OKAPI-BM25 and Bayesian interface network model shows that performance of the former is more efficient than the latter. Another point that needs to be highlighted here is that the proposed IR system managed to retrieve all related documents at level 60. However, it achieves 100% recall at position P60 as shown in Fig. 3. In contrast, other models achieved only 82% recall for the first 100 retrieve documents.

The analytic results demonstrated that the newly evaluation function achieves better effectiveness performance than that of OKAPI- BM25 and Bayesian interface network model. Therefore, hypothesis 1 is accepted.

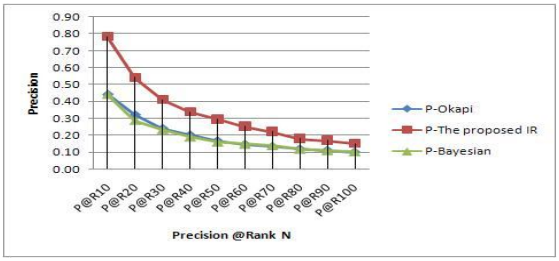

Fig. 2 Comparison of Precision for the three evaluation functions.

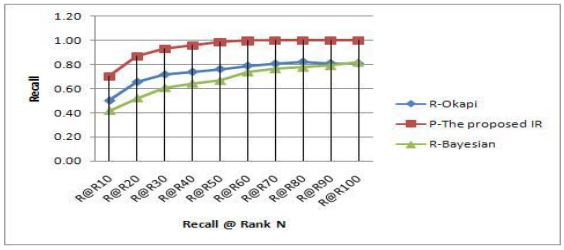

Fig. 3. Comparison of Recall for the three evaluation functions.

G. *The Analysis of Recall-Precision Measure*

The average of recall- precision relationship is examined in this section. Fig. 4 shows the performance of the proposed system whereas the precision value was 0.86 for only 10% of relevant documents. The precision values of the proposed system are varying between 0.83 and 0.91. 100% of precision value is achieved when system's recall was 100%. Moreover, only 9- 17% of retrieved documents are not relevant to user query and they appear in low rank until the system will retrieve all relevant documents. The achieved results are very near to user's expectation since he or she is looking to have all top ranked documents are relevant and most of the relevant documents appear at top position. In contrast, the precision value of Bayesian network interface model was 0.57 at recall of 10% and 0.7 when it reaches the maximum. However, 0.47 of precision value is achieved at the 100% recall. Therefore, more than half of the retrieved documents are not relevant in the results retrieved by Bayesian network interface model. On the other hand, the precision value of OKAPI-BM25 model was 0.66 at 10% recall and begins to decrease until it reaches the value of 0.48. The achieved results in the proposed system

ISBN: 978-988-19251-5-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

doesn't depend only on frequency of the terms within the document, it depends on the importance of the term based on HTML tag and on the position of the terms within the document. Therefore, based on the achieved results in this experiment, hypothesis 1 is also accepted.

H. Effectiveness of Indexing Schema

The essence of our approach in the proposed indexing schema was facilitating the required process of querying the information from the database and therefore this will increase the applicability of the proposed IR system to deal with large

volume of documents. i.e.: 8000 documents. Due to the nature of the Oracle engine is help to achieve our objective. For example, ranking results is done by just adding *"order by"* closure in SQL. In addition, document evaluation is done much faster than using old method because obtaining some factor (i.e., minimum term distance) is obtained by querying a simple SQL statement. Therefore, hypothesis 2 is accepted.

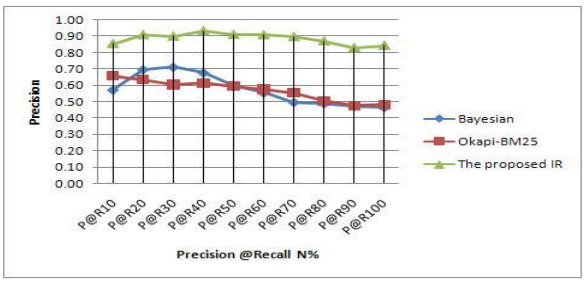

Fig. 4. Average recall-precision relationship for the three evaluation

VIII. Conclusion and Future Work

The current study contributes to retrieved relevant documents by applying new evaluation function and using new indexing schema. An experimental study was conducted to examine its performance against two other methods. However, the characteristics of the new evaluation function are:

- Evaluates the documents independently of other documents in the collection.
- Uses the probability of query keywords within the document.
- Uses the terms proximity concept.

The proposed system has been implemented and tested on dataset from Carnegie Mellon University. A comparison is done with two well known evaluation functions which are Okapi-BM25 and Bayesian network interface model. The findings demonstrated that the new evaluation function of the proposed system achieves better performance and shows improvement in the quality of retrieved documents than the mentioned methods.

In future research, we plan to apply the proposed system with large set of queries and large set of documents. This research also provides a foundation for a future study that will examine the impact of adjacent terms within a sentence and terms that appear in consecutive sentences to deal with them differently when the minimum distance is calculated.

References

[1] A. Al-Dallal and R.S. Abdul-Wahab, "Genetic Algorithm Based to Improve HTML Document Retrieval", Second International Conference on Developments in eSystems Engineering (DESE), 2009, Page(s): 343 – 348.

[2] A. Aly, "Applying Genetic Algorithm In Query Improvement Problem", *In: Information Technologies and Knowledge* Vol.1, 2007, p 309 – 316.

[3] B. Liu, "*Web Data Mining*", Springer-Verlag New York, LLC, Dec 2006, p204- 208.

[4] B. Minaei-Bidgoli and W. Punch, "Using genetic algorithms for data mining optimization in an educational web-based system". *In Genetic and Evolutionary Computation Conference*, Chicago, USA, 2003, pp. 2252–2263.

[5] C. Lopez-Pujalte, V.P. Guerrero-Bote, F. de Moya-Anegon, "Genetic algorithms in relevance feedback: a second test and new contributions", Information Processing & Management 39 pp (2003) 669–687.

[6] C. Tian, T. Tezuka, S. Oyama, K. Tajima, K. Tanaka, "Improving web retrieval precision based on semantic relationships and proximity of query keywords". In: Bressan, S., K ¨ ung, J., Wagner, R. (eds.) DEXA 2006. LNCS, vol. 4080, pp. 54–63. Springer, Heidelberg (2006).

[7] D. Vrajitoru, "Genetic Algorithms in Information Retrieval". *AIDRI97: Learning; From Natural Principles to Artificial Methods.* , Genève, June. 1997.

[8] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval". *Information Processing and Management*, 24(5):513-523, 1988.

[9] J. Carlberger, H. Dalianis, M. Hassel, and O. Knutsson, "Improving precision in information retrieval for Swedish using stemming". *In Proceedings of NODALIDA '01 - 13th Nordic conference on computational linguistics.* Uppsala, Sweden, 2001.

[10] M. H. Marghny and A. F. Ali, "Web Mining Based On Genetic Algorithm", *AIML 05 Conference*, 19-21 Dec 2005, CICC, Cairo, Egypt.

[11] P. Pathak, M. Gordon and W. Fan. "Effective information retrieval using genetic algorithms based matching functions adaption", *In: Proceedings of: 33rd Hawaii International Conference on Science (HICS),* Hawaii, USA, 2000.

[12] R. Cummins, C. O'Riordan, "Evolving local and global weighting schemes in information retrieval", *Information Retrieval***.** Boston. Vol. 9, Iss. 3, 2006, p. 311.

[13] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu,and M. Gatford. Okapi at TREC-4. In Harman, D. K., editor, *Proceedings of the Fourth Text Retrieval Conference*, pages 73–97. NIST Special Publication 500-236, 1996.

[14] S. Kim, B. Zhang, "Genetic Mining of HTML Structures for Effective Web-Document Retrieval", *Applied Intelligence*, v.18 n.3, p.243-256, 2003.

[15] S. Kim, B. Zhang, "Web-Document Retrieval by Genetic Learning of Importance Factors for HTML Tags". *In Proceedings of PRICAI Workshop on Text and Web Mining'2000.* pp.13-23.

[16] W. Fan , E. Fox , P. Pathak , H. Wu, "The effects of evaluation functions on genetic programming-based ranking discovery for Web search", Research Articles, Journal of the American Society for Information Science and Technology, v.55 n.7, May 2004, p.628-636.

[17] W. Song, S.C. Park. "Genetic algorithm for text clustering based on latent semantic indexing". Computers & Mathematics with Applications. Volume 57, Issues 11-12, June 2009, Pages 1901-1907.

[18] Carnegie Mellon University Data Set [Online]. Available: http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

[19] Description of LSI, [Online]. Available: http://en.wikipedia.org/wiki/Latent_semantic_indexing

[20] B. Klabbankoh, O, Pinngern "Applied Genetic Algorithms In Information Retrieval", Available: http://www.ils.unc.edu/~losee/gene1.pdf

[21] M. Persin, "Document Filtering for Fast Ranking", in SIGIR '94 Proceedings of the 17th annual international ACM SIGIR Springer-Verlag New York, 1994.

ISBN: 978-988-19251-5-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)