ISSN 2278-7763

Text-Based Information Retrieval System using Non-Linear Matching criteria

[(1)]Dr. Loay Edward George, [(2)]Ibrahem Amer Hameed

[(1)]College of Science, Computer Science Department, Baghadad University, Baghdad, Iraq; [(2)] College of Science, Computer Science Department, Baghadad University, Baghdad, Iraq
[(1)] Loayedward57@yahoo.com
[(2)] Ibrahem_star_2005@yahoo.com

ABSTRACT

The Web has a huge amount of information, which retrieved using information retrieval systems such as search engines, this paper presents an automated and intelligent information retrieval system that retrieves the HTML files and ranks them according to the degree of their similarities with the query HTML document. The most similar documents are top ranked. The developed system depends on the textual content of documents. The frequency of occurrence of each word, its printing attributes, and its critical position in the Web document were all used to determine the degree of significance of each word. The main challenge in the retrieval process is the matching task; the highly scored keywords may cause a biasing on the matching decision, so to handle this process, a non linear mapping function utilized. The test results indicated that the precision is increased about 5.6%, while the recall increased about 2.63%.

Keywords : Information retrieval, non-linear mapping, HTML documents, precision, recall, score, Web, Search engine, query.

1 INTRODUCTION

The World Wide Web, or simply the Web, serves as a huge, widely distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e-commerce, and many other information services [1],[2],[3]. The Web is a popular and interactive medium to collect, disseminate, and access an increasingly huge amount of information [1],[2].

The Web represents a new framework that is rather different with respect to the traditional IR framework and sheds new and difficult challenges [4],[5]. Web presents particular characteristics that limit the existing IR technologies and determines the need to design new information access technologies [6] among these characteristics:

- Web is possibly the biggest dynamic information resource.
- Web presents a structure of linked pages.
- Web is growing and updating at very high rate.
- Web is very heterogeneous.

Imprecision and vagueness characterize several tasks in Web IR, such as assessing the relevance of Web pages, dealing with the multimedia nature of information, identifying spam problem, discovering deception, etc. Furthermore, due to this complexity, any major advance in the field of information access on the Web requires the application of intelligent techniques [1], [5].

Soft Computing (SC) techniques constitute a synergy of methodologies (e.g. fuzzy logic, neural networks, probabilistic reasoning, rough-set theory, evolutionary computing and parts of machine learning theory). They are useful for solving problems requiring some form of intelligence [6]. The basis of SC is its tolerance to imprecision, uncertainty, partial truth, and approximation. Because of these properties SC can provide very powerful tools for modeling the activities related with the Web information access problem [7], [8].

2 TRADITIONAL INFORMATION RETRIEVAL SYSTEM

Information retrieval (IR) is searching for documents, for information within documents, metadata about documents, as well as for searching the existing relational databases in the World Wide Web[4,5,11]. There is overlap in the usage of the terms "data retrieval", "document retrieval", "information retrieval", and "text retrieval". Each has its own body of literature, theory, praxis, and technologies. IR is interdisciplinary science; it is based on computer science, mathematics, library science, information science, information architecture, cognitive psychology, linguistics, statistics, and physics[4,5]

Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications [6]. Figure (1) the typical layout of a classical information retrieval system.

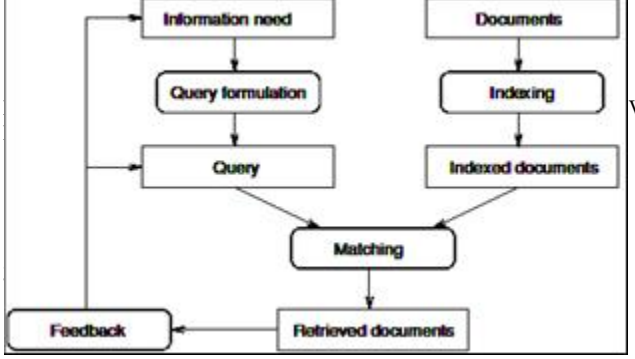

Fig. 1. Classical information retrieval system

3 Proposed Work

The system supports a client-side Web query to retrieve all similar Web documents according to their topic relevancy with the query document. At the enrollment stage, the lists (or records) of weighted keywords extracted from the Web pages are stored in the system's database. The URL address of each document is associated with its own list of keywords which is considered as the document descriptive information.

During the retrieval stage (i.e., after the enrollment phase) the queried Web is analyzed to extract its list of keywords. Then, the extracted list of keywords is matched with the pre-extracted and archived lists in the database; to decide the degree of relevancy between the queried Web document and the archived Web documents (in the database).

Since the Text Based Information Retrieval System TBIR is a sort of supervised classification process; it is composed of two phases:

(i) The enrollment (off-line) phase: it includes the operations conducted on server side only.

(ii) The retrieval (on-line) phase: some of its operations are conducted on client side, other operations accomplished at server side.

4 Enrollment

This phase is server perspective, because there is no client interaction during this phase. So, it is an offline process. The extracted list of keywords from each HTML document is stored in the database with the URL address of that Web HTML document. It involves the following:

4.1 Preprocessing

As any data mining application, preprocessing/data cleansing operation is an obligatory step, which involves the following:

1. Downloading and Conversion: It implies the conversion of UTF-8 code of the Web pages to the corresponding traditional computer representation of characters (i.e., the ASCII code).
2. Tokenization and Filtering: tokenization operation converts a sequence of characters stored in a string block to a finite set of understandable words stored in an array of strings.

4.2 Text Operations

The established TBIR works on a textual content data of the HTML documents, this imposes that text operations are inevitable, (namely, these operations belong to the text mining field). The TBIR has a text analyzer that analyses the extracted keywords by implementing certain textual mining operations. The implemented text operations are the stop words removal and the stemming. The stemming is the process of originating words to their roots[5],[6],[7],[8].

4.3 Weighting

The extracted words form the text operations are given weights according their frequency of occurrence and the styling features that the words had taken .

The calculation of the weight is performed using the following equations:

$$F_n(j) = \frac{F(j)}{\sum_{i=1}^{n} F(i)} \times \%100 \qquad (1)$$

Where – F(j) is the frequency of occurrence of the Jth keyword, – Fn (j) is the relative frequency, i.e. probability of occurrence, – N is the total number of extracted keywords.

$$w_f(j) = \sum_{i=1}^{8} w_i(j) \qquad (2)$$

Where – wi(j) is the corresponding weight of ith styling feature of the extracted jth word. The value of the weight wi(j) is determined using the mapping function:

$$w_i(j) = \begin{cases} \alpha_i & \text{if the } i^{th} \text{ feature of } j^{th} \text{ keyword is true} \\ 0 & \text{otherwise} \end{cases}$$

$$w_{Overall}(j) = F_n(j) \times w_f(j) \qquad (3)$$

Where, $w_{Overall}(j)$ – the overall weight value of the j_{th} keyword.

4.4 Keyword Sorting

The extracted keyword list has to be sorted in descending order according to the corresponding overall weight of each keyword in the list. After generating the weighted words they are

ISSN 2278-7763

stored in the system's database.

5 The Retrieval

This phase involves all previously mentioned operations in the enrollment phase (but without database storage step). All these operations are directly conducted at server side under on-line bases with users. Once the client starts the mining operation over a specific query document, the client side program sends a query request with the URL address of the queried HTML document; then the system, at the server side starts the on-line retrieval operation. This operation involves the following main tasks:

5.1 Preprocessing and Keyword Extraction

This stage holds all the operations mentioned in the enrollment phase but the resulted list of keywords is not stored in the database; rather it is used to be matched with the records already stored at the enrollment.

5.2 Scoring

In this stage, matches between the query keywords record and the keyword records stored in the database are performed. The matching is repeated for all registered records in the retrieval database. The result of each matching instance is named a matching score. The match score is determined as the sum of the overall weights of the common keywords found in both matched records. The matching score is calculated using the following equation:

$$S(m) = \sum_{k_w \in Q \cup DB_m} W_{Overall}(k_w | Q) W_{Overall}(k_w | DB_m) \quad (7)$$

Where, kw– a keyword found in both query and mth record of the database, Woverall (kw |Q) – is the overall weight of the keyword, kw , found in the query record, Woverall (kw |DBm) – is the overall weight of the keyword, kw , found in the mth record of the database, and S(m) – is the score of matching between both query record and mth record in the database.

The conducted tests have shown that the existence of few highly weighted common keywords may cause bias in grading results. To handle the problem of size dependency and effectiveness of highly weighted keywords, a non-linear mapping function (i.e., S- shaped function) was used to determine the matching score.

$$\mu(w_k) = 1 + \tanh(w_k) = \frac{1}{1 + e^{-2w_k}} \quad (4)$$

Where, wk– is the sum of the overall weights of the keyword (k) found in both the query and the database record.

So, instead of applying equation (5) the total matching score between the two records is determined using:

$$S(m) = \sum_{k_w \in Q \cup DB_m} \mu(W_{Overall}(k_w | Q) W_{Overall}(k_w | DB_m)) \quad (5)$$

5.3 Ranking

The ranking is considered the last process in the retrieval phase of TBIR. When the list of ranks (due to matching of query record with all database records) is established, it is sorted in descending order with the corresponding URL addresses. Then, the top listed URLs are tabulated in output text file sent as output data frame to the client side. total matching rank is determined using the following equation:

$$R = CW_{Overall} + B \quad (6)$$

Where, Woverall – is the determined matching score, B – is the bonus value due to the styling features of common keywords, C – is a constant used to adjust the significance of Woverall., R – is a total matching rank.

6 Performace Measures

The measures usually employed to evaluate the performance of an information retrieval system are precision and recall. Precision (Pr) is "How many of the retrieved results are relevant" and the recall (Re) is "How many of the relevant results are retrieved"; mathematically, they defined as [9],[10],[11],[12],[13].

$$Pr = \frac{R}{N} \quad (8)$$

$$Re = \frac{R}{M} \quad (9)$$

Where, R– is the total number of the relevant retrieved items, N – is the retrieved items, M– is the relevant items.

7 Expermental Results

The total number of HTML documents used in the testing stage is 420 documents, which belong to 9 classes. Those documents are preprocessed and their produced records are deposited in the database. Table (1) shows the topic classes and the number of used HTML documents belonging to each class.

TABLE 1
Classes of subjects Used for TBIR

ISSN 2278-7763

#	Class	No. of HTML Documents
1	Astronomy	40
2	Chemistry	40
3	Computer Science	60
4	Modern History	45
5	Ancient History	45
6	Physics	60
7	Sports	40
8	Medicine	45
9	Mathematics	45
Total		420

As query samples, three documents are taken from each of the nine classes, so as a total number of 27 documents is used as a query samples. The conducted test was aimed to measure the precision, recall and the time required to execute each query.

To measure the effect of using non-linear function (i.e., tanh(x)) is investigated in terms of precision, recall and time, The results show that the use of s-mapping function has enhanced the system accuracy performance.

8 Conclusions

The World Wide Web, serves as a huge, widely distributed, global information service center and has a huge amount of information, those information are retrieved to the user on demand using information retrieval systems such as search engines, this paper presents an automated and intelligent information retrieval system that retrieves the HTML files and ranks them according to the degree of their similarities with the query HTML document. The most similar documents are top ranked in the fetched list of addresses. The developed information retrieval system depends only on the textual content of Web documents. The frequency of occurrence of each extracted word, its printing attributes, and its critical position in the Web document were all used together to determine the degree of significance of each extracted word.

The proposed method for Web document retrieval based on non-linear mapping that improved the precision about 5.6% rather than using the traditional hard computing. Table (2) states the results

No.	Class	Linear Mapping		Non-Linear	
		Prec.	Recall	Prec.	Recall
1	Astronomy	1	0.75	0.8	0.6
2		1	0.75	1	0.65
3		1	0.75	0.97	0.7
4	Chemistry	1	0.66	0.8	0.4
5		0.8	0.53	0.8	0.4
6		1	0.66	0.87	0.5
7	CS	0.93	0.48	0.8	0.4
8		0.7	0.36	0.77	0.38
9		0.93	0.48	0.73	0.37
10	ModernH.	0.97	0.72	0.73	0.55
11		0.9	0.68	0.9	0.68
12		1	0.67	0.67	1
13	Math	0.83	0.53	0.93	0.62
14		0.73	0.49	0.86	0.58
15		0.97	0.65	0.93	0.62
16	Medicine	1	0.75	0.97	0.72
17		1	0.75	0.97	0.72
18		1	0.75	1	0.75
22	AncientH.	1	0.66	1	0.66
23		0.87	0.57	0.87	0.57
24		0.93	0.62	0.9	0.6
25	Physics	1	0.5	0.97	0.48
26		0.93	0.57	0.93	0.57
27		1	0.5	1	0.5
28	Soccer	1	0.67	0.97	0.64
29		1	0.67	0.93	0.6
30		1	0.67	0.97	0.64
		0.94	0.62	0.89	0.59

TABLE 2
The effect of s-mapping function

References

[1] Chu, H., "Information Representation and Retrieval in the Digital Age", American Society for Information Science and Technology, ISBN 1-57387-172-9, Vol. 9, Pp. 111-112, 2003.

[2] Adam, S., Horst B., Mark L., Abraham K., "Graph-Theoretic Techniques for Web Content Mining", Springer, ISBN 981-256-339-3, 2005.

[3] Anthony, S., "Web Mining Applications and Techniques", Idea group publishing, ISBN 1-59140-414-2 -- ISBN 1-59140-415-0 (ppb) -- ISBN 1-59140-416-9 (E-book), 2005.

[4] Baeza-Yates, R., & Ribeiro-Neto, B., "Modern Information Retrieval", ISBN 0-201-39829-X, Addison Wesley, 1999.

[5] Christopher, D., M., "An Introduction to Information Retrieval", Springer, ISBN-13 978-3-540-31588-9, 2009.

[6] Djoerd, H., "Information Retrieval Models", J. Information Retrieval, John Wiley and Sons, Ltd., Vol. 29, Pp.1-24, 2009.

[7] William, R., H., Diane, L., E., David, H., H., Stephanie L., W., "Towards New Measures of Information Retrieval Evaluation", SIGIR '95 Proceedings of the 18th annual international, ACM SIGIR conference on Research and development

in information retrieval, ISBN:0-89791-714-6, Pp. 164-170 , 1995.

[8] Subhendu, K., P., Deepak, M., Bikram, K., R., "Integration of Web mining and Web crawler: Relevance and State of Art", (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, Pp. 772-776, ISSN : 0975-3397 772, 2010.

[9] G.S. Tomar,Shekhar Verma, Ashish Jha, "Web Page Classification using Modified NaïveBayesian Approach",2006.

[10] Pasca, M. and Harabagiu, S., "High Performance Question/Answering". In Proceedings of the 24th International Conference on Research and Development in Information Retrieval, Pp. 366–374, 2001.

[11] Salton, G., and Buckley C., "Term-Weighting Approaches in Automatic Text Retrieval", Information Processing and Management, Vol. 24, No. 5, Pp. 513–523, 1988.

[12] Daniel T.Larose Discovering Knowledge in Data an Introduction to Data Mining, Wiley, corporation,pp.90-107 ,2005.

[13] Turtle, H., "Inference Networks for Document Retrieval", Ph.D thesis, Department of Computer Science, University of Massachusetts, Amherst, MA 01003, 1990. Available as COINS Technical Report 90-92.

