ESKİŞEHİR TECHNICAL UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY
A- APPLIED SCIENCES AND ENGINEERING

2020, 21(1), pp. 182 - 198, DOI: 10.18038/estubtda.615103

ON THE USEFULNESS OF HTML META ELEMENTS FOR WEB RETRIEVAL

Ahmet ARSLAN[1,*]

[1] Computer Engineering Department, Faculty of Engineering, Eskişehir Technical University, Eskişehir, Turkey

ABSTRACT

Web retrieval studies have mostly used URL, title, body, and anchor text fields to represent Web documents. On the other hand, HTML standards provide a rich set of elements to define different parts of a Web page. For example, meta elements are used to provide structured metadata about a Web page not to end users, but instead to browsers or crawlers. However, it is unclear whether meta tags are or are not useful for Web retrieval, as most of the previous studies leveraged URL, title, body, and anchor text fields. In this work, we examine the usefulness of two meta tags, namely keywords and description, based on ad-hoc tasks of previous TREC studies. Through experiments on the standard TREC Web datasets and several query sets, our results using the state-of-the-art term-weighting models show that the utilization of description field systematically increases the retrieval effectiveness, to a statistically significant degree most of the time. By contrast, the employment of keywords field may cause a significant deterioration in retrieval effectiveness for certain term-weighting models.

Keywords: Information Retrieval, Web Retrieval, Meta Tags, ClueWeb, HTML

1. INTRODUCTION

Text Retrieval Conference (TREC) Web datasets (e.g. GOV2 and ClueWeb), which are widely used for retrieval experiments, are the collections of Web pages crawled from the Internet. Therefore, these datasets are composed of documents written in Hyper Text Markup Language (HTML). It is possible to derive several structured document representations from a HTML file, such as body, title, headings, and paragraphs. Furthermore, each HTML page on the Internet is accessed by its Uniform Resource Locator (URL). Apart from the content of the HTML document (e.g. body and title), another source of information is the document's URL.

Two pages in the Web can be connected to each other by means of a hyperlink. The source document is the one containing the hyperlink whereas the target document is the one the hyperlink points to. In that sense, the Web can be thought of a directed graph in which pages are the edges and hyperlinks are the vertexes. Furthermore, a hyperlink has clickable text associated to it, which is referred to as anchor text. The hyperlinked structure of the Web gives rise to another field "anchor texts from *incoming* hyperlinks". It is important to emphasize that anchor texts are extracted from incoming links, not outgoing links. Thus, the anchor texts are not written by the author of the Web page but instead by a number of people, which makes the manipulation of anchor text harder to achieve. Robertson et al. [1] refer to anchor text as *repeatable field* to emphasize the multivalued nature of this field. Recall that there can be multiple hyperlinks that point to a page and the same keyword might occur multiple times in anchor texts.

Different parts of a document (e.g., title, body, URL, and anchor) are called document representations or fields in the IR literature. Hereafter, the terms 'fields' and 'document representations' are used interchangeably. The standard practice in the Web retrieval literature is to work on the aforementioned four fields: title, body, URL, and anchor text [2]. However, there is another set of fields that can be extracted from the `<meta>` tag, which provides metadata about the HTML document [3].

*Corresponding Author: aarslan2@eskisehir.edu.tr
Received: 03.09.2019 Published:31.03.2020

Meta elements allow to specify page description, keywords, author, last modified date and refresh rate of the page as well as other metadata. The most interesting property of the metadata is that they are not rendered by a Web browser (i.e. not displayed on the page), but can be used by Internet browsers, crawlers of search engines, or other Web services. Although considerable research has been devoted to URL and anchor text, rather less attention has been paid to this invisible content. In the present work, we focus on two meta elements, namely keywords and description.

In spite of the fact that a few participants of the earlier TREC Web tracks (ran through 2001 to 2004) used the keywords and description meta tags, the main focus of these studies is not on the impact of the keywords and description fields. To our knowledge, no previous work has analyzed the usefulness of keywords or description fields on the latest TREC Web tracks series (ran through 2009-2014), which use the new ClueWeb datasets.

1.1. Research Objective and Contributions

The main objective of the present paper is to analyze the usefulness of keywords and description meta tags in Web retrieval effectiveness. We base our analysis on the standard TREC Web datasets (GOV2, ClueWeb09 and ClueWeb12) and the associated query sets released between the years 2004 and 2018.

In this work, we make the following contributions:

(1) Experiments to determine which meta tag (description, keywords, or combination of both) is the most useful in improving Web retrieval effectiveness in a static retrieval setting.
(2) Present a preliminary discussion of the possibility of a selective application of keywords and description fields on a per-query basis.

The rest of this article is organized as follows. The fundamental HTML elements as well as meta tags are described in the next section in order to provide a context for this study. In Section 3, we discuss previous work. Section 4 describes our experimental setup, the benchmark collections and retrieval methods used in the analysis. Section 5 compares and discusses the retrieval effectiveness obtained for different document representation combinations (with and without meta elements). Section 6 compares the retrieval effectiveness of individual document representations and discusses the possibility of a selective retrieval approach that predicts the best representation on a per-query basis. Section 7 concludes the article with directions for future work.

2. HTML, META ELEMENTS AND THE CHALLENGES OF WEB RETRIEVAL

HTML is the standard markup language for Web pages. HTML is a hierarchal structure that begins with a `<html>` tag, usually contains a `<head>` and `<body>` tag, and elements can be nested within elements. Figure 1 shows HTML source code of a typical Web page and how it is rendered in a Web browser. Figure 1 (a) illustrates the basic components (title, body, hyperlink, heading, meta tags, and comment) of a simple HTML file. The parsed DOM (Document Object Model, see http://www.w3.org/DOM) tree shown in Fig. 1 (b) demonstrates the hierarchal structure of HTML, which usually contains a `<head>` and `<body>` tag, and elements can be nested within elements. Figure 1. (c) shows the rendered version of the HTML document. Note that the meta elements and the comment text are not displayed in the rendered version.

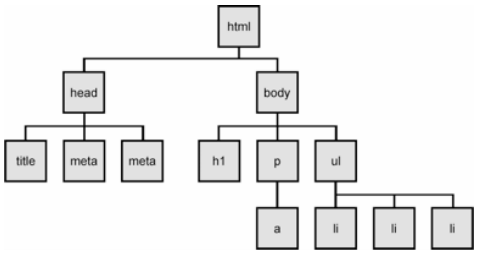
2.1. The Title and Body

The `<title>` and the `<body>` tags define the title and the body of the document, respectively. The body of the document contains all the contents of an HTML page, such as main text, paragraphs, headings,

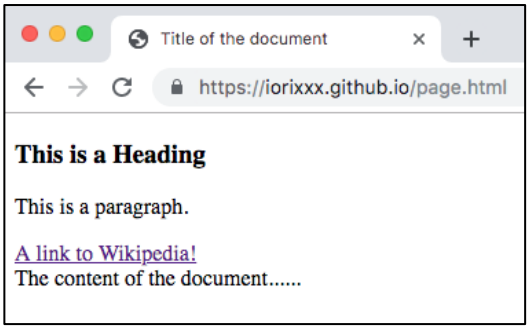
tables, images, hyperlinks, etc. The standard practice to represent a Web document using a single field is to merge title and body fields into one.

```
<!DOCTYPE html>
<html>
<head>
<title>Title of the document</title>

  <meta charset="UTF-8">
  <meta name="description" content="Free Web
tutorials">
  <meta name="keywords"
content="HTML,CSS,XML,JavaScript">
  <meta name="author" content="John Doe">

</head>

<body>

<h3>This is a Heading</h3>
<p>This is a paragraph.</p>

<a href="https://www.wikipedia.org/">A link to
Wikipedia!</a>

<!-- This is a comment -->
<br>

The content of the document......

</body>
</html>
```

(a) HTML Code



(b) DOM Tree



(c) Rendered Page

Figure 1. A mock-up of a typical Web page: (a) The source code of the HTML page; (b) The DOM tree of the page; (c) How the HTML page is rendered by a modern Web browser.

2.2. The Meta Elements

The `<meta>` tag contains metadata about the HTML document. Typical metadata are page description, keywords, author of the document, last modified, and refresh rate. Metadata are not displayed on the page but can be used by browsers, search engines, or other Web services. The keywords and description meta tags are two of the most probable candidates to be used by search engines for retrieval purposes. The keywords meta tag contains a comma-separated list of terms relevant to the current page. The description meta tag describes the content of the document. There can be only one description element per document.

2.3. The Headings and Paragraphs

It is possible to define headings with different importance levels using the `<h1>` to `<h6>` tags in which `<h1>` is the most important and `<h6>` is the least important. The `<p>` element represents a paragraph which can be used to group related content. Some space is added before and after a paragraph by browsers in order to separate it from adjacent blocks.

2.4. The URL

Each Web page has a unique URL (also called Web address), which is used to access that page. The URL of the mock-up page, http://iorixxx.github.io/page.html, can be seen in the address bar of the browser in Fig. 1(c). Since users may want to type a Web address, human readable and memorable URLs are desirable. The URL and anchor fields are the two sources of information that cannot be obtained from the content of the document (i.e. cannot be seen in Fig. 1(a)). It is important to note that

the underlined text: "A link to Wikipedia!" in Fig. 1(c) is not the anchor text of the page displayed in Fig. 1(a), but it is the anchor text of the Web page with the URL https://www.wikipedia.com.

2.5. The Anchor Text

The HTML `<a>` tag is used to create a hyperlink between two documents. When the user clicks the link on the source document, browser navigates to the target document. The document shown in Fig. 1(a) is the source document. The `href` attribute specifies target document, which is https://www.wikipedia.com in Fig. 1(a). Anchor text is the text associated with a link in a source document. The anchor text document representation used in IR includes the concatenation of all anchor texts obtained from incoming links (i.e. source documents), which is assumed to describe the target document. Since, anchor text extraction requires non-trivial computation (i.e. traversal of the whole graph), anchor texts of the datasets are usually precompiled and released as supplementary data. For example, anchor text data for both the ClueWeb09[1] and ClueWeb12[2] datasets are provided by Djoerd Hiemstra [4]. By contrast, anchor text is not publicly available for the GOV2 dataset.

2.6. The Challenges of the Web and HTML Parsing

The idiosyncratic properties of the Web bring many challenges to the IR research. Parsing the HTML has proven to be difficult. This challenge is recognized by the organizers of the NII Testbeds and Community for Information access Research (NTCIR) who draw attention to the difficulty of handling the raw HTML content by providing the extracted content with professional tools of Sogou.com [5]. On this account, Brin and Page [6] state that "Any parser which is designed to run on the entire Web must handle a huge array of possible errors. These range from typos in HTML tags to kilobytes of zeroes in the middle of a tag, non-ASCII characters, HTML tags nested hundreds deep, and a great variety of other errors that challenge anyone's imagination to come up with equally creative ones."

The is a number of open source libraries for HTML parsing, among which the JSoup is the one that is actively developed and used in popular open source IR toolkits such as Terrier [7] and Anserini [8]. The JSoup library provides a single convenience method named `org.jsoup.nodes.Element#text()`, to obtain the text content of a given HTML document. The method roughly returns the concatenation of the body and title of the input HTML. The text extracted from the example HTML shown in Fig. 1 is "*Title of the document This is a Heading This is a paragraph. A link to Wikipedia! The content of the document.....*" Note that the text freed from HTML tags is the equivalent of the visible text in Fig. 1(c) in which the meta elements and the comment text are not included.

Furthermore, as illustrated in Fig. 1, HTML source code looks quite different from its rendered version. The end users interact with the rendered version in a Web browser whose primary function is to request HTML source, CSS, JavaScript and images from a server and render them based on Web standards and specifications. JavaScript can be used to display text on a page, which does not exist in the source code of the page. By contrast, CSS can be used to print invisible text (in the same color as the background of the Web page) on a page that exist in the source code. Furthermore, unwanted popup windows can be triggered via JavaScript. To the best of our knowledge, the aforementioned aspects have not been yet investigated for the TREC Web datasets. This may be due to the fact that HTML source alone is not enough to obtain a correct rendered version of a page. TREC organizers noted that ClueWeb09's additional files can be supplied if requested by a researcher. Furthermore, there is a page rendering

[1] https://djoerdhiemstra.com/2010/anchor-text-for-clueweb09-category-a/

[2] http://wwwhome.ewi.utwente.nl/~hiemstra/2013/anchor-text-for-clueweb12.html

service[3] for the ClueWeb12 dataset, in which you can see the rendered version of a page for a given TREC document identifier.

But even in the official TREC assessment process, some pages may not render properly. That is why a brand new relevance level (-3) is introduced in the TREC 2016 Tasks Track entirely dedicated to the pages that are not rendered properly at the time of evaluation [9].

3. RELATED WORK

The URL, title, body, and anchor fields are the most common representations of Web documents used in the Information Retrieval (IR) literature [2]. The usefulness of anchor text for Web retrieval was recognized early on, dating back to 1994 [10]. The anchor text has been extensively studied since then [11] [12] [13] [14]. For instance, Anh and Moffat [15] investigated the role of anchor text in ClueWeb09 retrieval; Ounis et al. [16] experimented with the sampling with and without anchor text in the learning-to-rank settings. The Google search engine used anchor text from its early development [6]. Anchor text seen as a brief summary of the page. It is sometimes argued that anchor texts from incoming links represent the document better than its actual content [17]. The reason is because the anchor texts are not written by the author of the document but by a number of other people, making manipulation of that text harder to achieve. In general, anchor texts are found to be useful for navigational queries in which the user wants to find a specific website and the query is the name of that website. In this scenario, the user searches for the query "youtube" to find the YouTube site instead of typing the full URL into a browser's navigation bar. This special type of search intent is referred to as "home page finding task" or "named page finding task" in the previous TREC Web track series ran through 2001 to 2004.

URL, another source of information, is also investigated for the home page and named page finding tasks. The URLs of documents are used as both query-dependent evidence (i.e. query terms are matched against the terms extracted form URLs) and query-independent evidence (i.e. the number of slashes ('/') in the URL). However, a special tokenization strategy may be required to ensure matching terms in a URL since URLs often contain abbreviations or glued terms. For example, Song et al. [18] use maximum prefix string matching for the task whereas Ogilvie and Callan [19] treat both the URL and query terms as sequence of characters and then compute a character-based *n*-gram generative probability. For the query-independent evidence, Westerveld et al. [20] classified URLs into four categories and estimated prior probabilities of being an home page on the basis of the URL type. The authors note that the number of slashes ('/') in a home page URL tends to be relatively small. Put differently, the URL of a home page tends to be higher in a server's document tree than other Web pages.

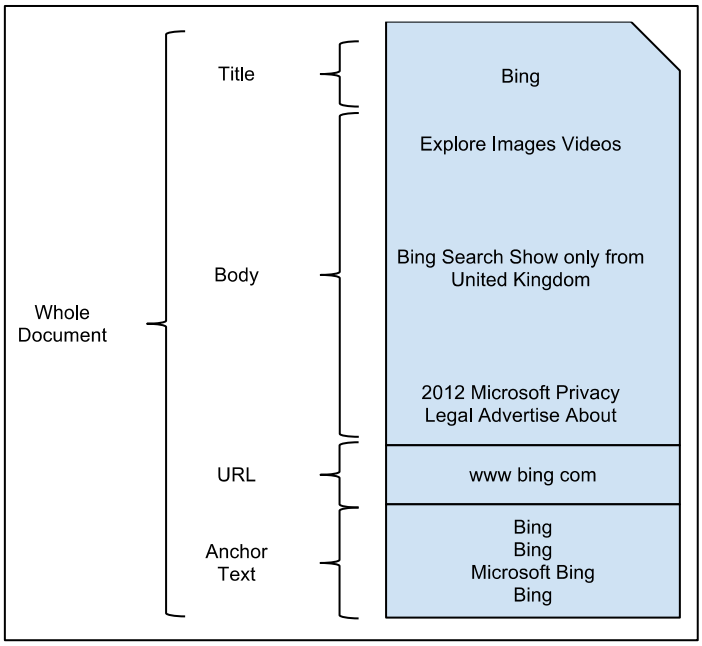
The `<title>` field is found to be repeated in child pages of a single website, or the default title "Untitled" is used in many pages [18]. Chibane and Doan [21] investigated headings `<h1>` to `<h6>` for the task of partitioning Web pages into coherent semantic bocks that correspond to a sequence of sub topical passages. Exploiting document structure by using multiple evidences (anchor, URL, link information, title, headings, etc.) was the general trend observed in the TREC 2004 Web Track [22].

In the studies focusing on the other components of IR process than document representation (i.e. a single document representation is used), anchor text is sometimes appended to the content in order to obtain a single field. For example, Zheng and Callan [23] treated anchor texts from incoming links as part of the document for three TREC Web datasets. However, Robertson et al. [1] warn against such practice as the volume of anchor text may *swamp* the remainder of the document in the existence of a large number of incoming links. For example, the number of incoming links in the anchor text data generated for the ClueWeb datasets varies between 1 and 10,000. Here, the maximum number of incoming links is not

[3] http://boston.lti.cs.cmu.edu/Services/clueweb12_batch/

coincidentally being 10,000. The anchor text extraction is based on an iterative algorithm in which the iteration number is set to 10,000. In other words, a maximum of 10,000 incoming links are considered.

The application of machine learning techniques to the ranking problem of IR has emerged recently. This new research area is called learning to rank (LETOR) [24] [16] [25]. Figure 2 illustrates how the URL, title, body, and anchor fields are utilized as a learning feature in the learning-to-rank settings.



(a) Different fields of a Web page.

ID	Feature Description
…	…
11	TF×IDF of body
12	TF×IDF of anchor
13	TF×IDF of title
14	TF×IDF of URL
15	TF×IDF of whole document
…	…
21	BM25 of body
22	BM25 of anchor
23	BM25 of title
24	BM25 of URL
25	BM25 of whole document
…	…
31	LM.DIR of body
32	LM.DIR of anchor
33	LM.DIR of title
34	LM.DIR of URL
35	LM.DIR of whole document
…	…

(b) LETOR v3.0 features.

Figure 2. Different parts of a Web page and learning-to-rank features derived from them: (a) Different document representations (fields) of a Web document, taken from [25]; (b) Selection of term-weighting model features for the LETOR v3.0 dataset, reproduced from [24].

3.1 Meta Tags

In this subsection, we briefly review the approaches of the groups that participated in the TREC Web track competitions (from 2001 to 2004) and utilized keywords and description meta tags as well as the previous studies that mention meta tags in different contexts than IR perspective.

To the best of our knowledge, the group of Carnegie Mellon University (CMU) [26] [19] is the first to explicitly use meta tags (keywords and description) along with other document representations including (*i*) full document text, (*ii*) title text, (*iii*) image alternate text, (*iv*) character trigram on URL, and (*v*) large font text (including headings). The CMU group forms language models from each of these document representations. The authors then combine the language models using a linear interpolation to form a new language model. When the effectiveness scores of individual document representations are compared, the representation based on meta tags proved to be the worst. However, the combination of all document representations within the generative language modeling framework is found to be much more effective than any of the individual document representations. The authors argue that the use of language modeling provides an effective mechanism for combining different document representations.

Later on, the members of the CMU group, Ogilvie and Callan [27], show that their generative language modeling framework attain better results than meta-search algorithms. The main difference between their language modeling approach and the meta-search techniques is that language modeling approach combines the document representations on the query term level, rather than as a post-retrieval score

combination. The authors also demonstrate that document representations that perform poorly (e.g. meta tags) can be combined with other representations to improve the overall effectiveness of the system.

Savoy and Rasolofo [28] used a document representation that accounts for page content along with its <title> and <meta> tags ("keywords" and "description") plus all anchor texts extracted from the other pages that link to that page. Song et al. [18] declare that they use anchor, title, meta and body fields to obtain a "whole page" representation. Similarly, Zhou et al. [29] construct a "pseudo-text" document representation that includes URL, title, anchor texts, and meta data. Tomlinson [30] use weighted matches in the title, URL, the first heading and some meta tags. However, none of the authors except Savoy and Rasolofo [28] report what the utilized meta field/data/tag is comprised of. Wen at al. [31] express that they extract keywords meta element, but do not reveal the details of how the keywords field is utilized in their ranking algorithm.

Recently, Roy et al. [32] investigated whether HTML tags should be stripped or not. The authors created two different indices (full and clean) of the ClueWeb09B and GOV2 datasets. The meta elements are implicitly included in the full index in which HTML tags are not stripped. Their results show that clean index, which is freed from HTML tags, produces better effectiveness results for the Language Model with Dirichlet prior smoothing (LM.DIR) and BM25. More recently, Gadge and Bhirud [33] divide Web documents into five layers (title, header, hyperlink, meta tag and body) with different priorities and propose a layered vector space model that can effectively combine these layers.

Although keywords and description fields were used by some of the previous TREC participants, their main focus was not on the keywords and description meta tags. To our knowledge, no previous work has extensively analyzed the usefulness of keywords or description fields using the new ClueWeb09 and ClueWeb12 datasets.

The meta tags are mentioned or discussed conceptually/theoretically in a number of previous studies, in which no IR experiments are performed (i.e. they do not report retrieval effectiveness). Regarding commercial search engines, Brin and Page [6] note that "metadata efforts have largely failed with Web search engines, because any text on the page which is not directly represented to the user is abused to manipulate search engines." Spirin and Han [34] argue that "the placement of spam content in meta tags might be very prospective from spammer point of view. Because of the heavy spamming, nowadays search engines give a low priority to meta tags or even ignore it completely." Lewandowski [35] lists meta tags as a query-dependent ranking factor by referring to search terms appearing in meta information such as keywords or description.

Another line of the previous research on meta tags (keywords and description) investigated commercial search engines' behavior to the meta tags. Turner and Brackbill [3] evaluate the effectiveness of using the <meta> tag to improve the retrieval of Web documents through Internet search engines. The authors suggest that the use of the keywords attribute substantially improves accessibility while the use of the description attribute alone does not. Craven [36, 37] analyze the variations in use of both keywords [37] and description [36] meta tags by Web pages in different languages. Zhang and Jastram [38] reveal that the "keywords" and "description" were the most popular single metadata elements used in the Web. Alimohammadi [39] give a good overview on the concept of the meta-tag in order to determine its applicability in indexing current documents of the Web.

4. EXPERIMANTAL METHOGOLOGY

This section describes the methodology we adopted for the experimental evaluation of the usefulness of description and keywords fields. We describe the datasets along with the corresponding query sets, the term-weighting models, the retrieval engine and different document representation combinations used in the analysis.

4.1. Datasets

To conduct experiments, we use three of the largest and newest collections of Web pages released by TREC, namely GOV2, ClueWeb09 and ClueWeb12. The ClueWeb09[4] and ClueWeb12[5] datasets contain 503,903,810 and 733,019,372 English Web pages, respectively. In this work, we use the "Category B" subsets of the full ClueWeb datasets. Table 1 lists the details of the three collections used in our experiments. In Table 1, '# Documents' columns show the number of documents that contain at least one term for each field. As it can be observed from Table 1, not all documents necessarily have all fields. This is even true for body and title fields (i.e. certain documents do not have a body tag). Missing fields is a challenge to the field-based Web retrieval. Although HTML standards are defined by the World Wide Web Consortium[6] and/or the Web Hypertext Application Technology Working Group[7], the Web seems to be chaotic and to lack structure. Furthermore, URL and anchor text data are not available for the GOV2 dataset.

Table 1. Statistics of datasets.

	GOV2		**ClueWeb09B**		**ClueWeb12B**	
	# Documents	**Avg. Field Length**	**# Documents**	**Avg. Field Length**	**# Documents**	**Avg. Field Length**
URL	N/A	N/A	50,219,202	7	52,342,942	9
body	24,775,882	875	50,169,043	780	51,956,294	716
title	21,843,524	6	49,687,322	7	50,818,446	7
anchor	N/A	N/A	44,392,147	413	29,808,461	47
keywords	3,274,724	22	29,690,318	27	22,479,343	26
description	2,653,661	15	24,802,592	22	26,280,591	23

Keywords field consists of 24.5 words on the average. This number is a little bit greater than one would expect, suggesting that keywords might not be assigned manually. It is worthwhile to note that plugins for automatically adding keywords and description meta tags exists for content management systems (e.g., WordPress, Joomla, and Drupal) that can be used to create websites. Thus, it might be possible that some of the keywords and description tags are automatically generated.

The average field length for anchor text reported in Table 1 must not be interpreted as the average number of words in all anchor texts. But instead, it is the average field length when all anchor text from incoming links (varying from 1 to 10,000 links) are merged into a single anchor field. A maximum of 10,000 incoming links are considered during the compilation of anchor text data for the ClueWeb datasets. Furthermore, 4,000,987 out of 44,392,147 documents are coming from the English Wikipedia for the ClueWeb09 anchor text. The number of documents that have at least one word in anchor field for the ClueWeb09 dataset is 32% greater than that of the ClueWeb12 dataset, which indicates that the ClueWeb09 dataset is more hyperlinked than the ClueWeb12 dataset.

For URL tokenization, we split URLs at non-letter characters and convert surviving tokens to lower case. URL and title are short fields, while, in contrast, body and anchor are long fields. As we can see, the fields listed in Table 1 have different average lengths and characteristics.

The usage statistics of meta elements in the TREC Web datasets are given in Table 2. The most frequently used meta tags across datasets are found to be keywords, description, generator, robots, and author. The usage of description and keywords elements are more prevalent in the new ClueWeb datasets than the GOV2 dataset.

[4] https://lemurproject.org/clueweb09
[5] https://lemurproject.org/clueweb12
[6] https://www.w3.org
[7] https://whatwg.org

Table 2. The top 10 most frequently used meta elements.

GOV2		ClueWeb09B		ClueWeb12B	
Meta Tag	**% Documents**	**Meta Tag**	**% Documents**	**Meta Tag**	**% Documents**
keywords	18	keywords	63	description	57
description	16	description	53	keywords	47
generator	10	generator	29	robots	21
author	5	robots	20	generator	16
robots	5	author	11	title	14
date	4	verify-v1	7	copyright	14
dc.title	3	copyright	7	google-site-verification	13
dc.publisher	3	revisit-after	7	author	11
dc.date	3	distribution	5	classification	9
progid	3	rating	4	if:show	4

4.2. Query Sets

Salient statistics of the query sets associated with the GOV2, ClueWeb09 and ClueWeb12 datasets are given in Table 3. The GOV2[8] dataset contains 25,205,179 Web pages crawled from U.S. government websites in the `.gov` and `.us` domains during early 2004 [40]. The dataset was used in the Terabyte Track series that ran through 2004 to 2006 and the Million Query Tracks that ran through 2007 to 2008.

The Million Query track series investigate whether shallowly judged many queries are more suitable to compare retrieval systems than using a deeply judged fewer queries. That is why the queries of Million Query tracks have relatively less relevant documents on the average than other tracks.

The ClueWeb09 [41] dataset and its successor, the ClueWeb12 [42] dataset, are used in six Web Tracks of TREC ran through 2009 to 2014. The organizers of TREC decided to end the Web Track in 2014 and launched a new one named "Tasks Track" which ran through 2015 to 2017 [9].

Table 3. Statistics of query sets

Dataset	**Track**	**Year(s)**	**# Queries**	**Abbreviation**	**Average query length**	**Average # relevant documents per query**
GOV2	Terabyte	2004-2006	149	GOV2	3.1	180.7 (±149.2)
	Million Query	2007	1524	MQ07	3.8	12.3 (±9.6)
	Million Query	2008	564	MQ08	5.2	5.2 (±6.7)
ClueWeb09	Million Query	2009	562	MQ09	2.6	15.5 (±25.7)
	Web	2009-2012	197	CW09B	2.5	95.3 (±74.0)
ClueWeb12	Web	2013-2014	100	CW12B	3.5	72.1 (±61.8)
	Tasks	2015-2016	85			
	We Want Web	2017-2018	180	NTCIR	2.5	160.3 (±91.0)

Although the tasks track focuses on the usefulness of a system in helping the user to complete the actual task that led the user to issue the query, it is possible to convert tasks tracks' relevance judgments to that of ad-hoc retrieval. Our query set does not include the Tasks Track 2017 because no team submitted runs for the document-based tasks; hence, no document judgments were constructed by TREC. After the discontinuation of the TREC Web Track, another IR forum named NTCIR continued ad-hoc Web retrieval by launching the "We Want Web" track, which includes an English subtask based on the ClueWeb12B dataset. The organizers of the NTCIR announce that ad-hoc Web search, in particular, is still of utmost practical importance and "We Want Web" track series will be continued for at least three years [5, 43].

[8] http://ir.dcs.gla.ac.uk/test_collections/gov2-summary.htm

4.3. Term-Weighting Models

Term-weighting models are used to rank documents based on a query hence they are the core component of retrieval systems. Therefore, various term-weighting models have been proposed for IR. Table 4 lists the term-weighting models used in our experiments, which contain a representative from the well-known retrieval families.

Table 4. Term-weighting models.

Model	Family	Reference
BM25	Probabilistic	Robertson and Zaragoza [44]
LM.DIR	Language model	Zhai and Lafferty [45]
DFI	Nonparametric	Kocabaş et al. [46]
LGD	Information based	Clinchant and Gaussier [47]

Except the Divergence From Independence (DFI), a parameter-free model, the other models have one or two hyper parameters. We use the default values taken from the Terrier [7] retrieval platform: LGD (c=1), LM.DIR (μ=2500), BM25 (b=0.75, k_1=1.2).

4.4. Experimental Settings

The Apache Lucene [48], whose usage in academic work has been gaining momentum [49], is used for indexing and searching in our experiments. We adopt the implementations of the term-weighting models from Terrier [7] (version 4.0) retrieval platform to Apache Lucene [50] (version 7.7.1). We keep the preprocessing of documents and queries minimum: after case-folding, we apply the Krovetz stemmer [51] and do not perform stop word removal.

We use the JSoup[9] library (version 1.11.3) to extract plain text from HTML files. JSoup throws a parsing exception or returns an empty string for some documents, which are skipped during indexing. We also discard the queries that do not have any relevant documents in the judgment set. The number of queries used in this study is 3,361 whose descriptive statistics are given in Table 3. We use the gdeval.pl[10] (version 1.3) evaluation script to calculate retrieval effectiveness scores. The relevance judgment sets of the Million Query [52] tracks are distributed in a different format than the remaining query sets; therefore, statAP_MQ_eval_v4.pl[11] evaluation script has to be used to calculate retrieval effectiveness scores for the Million Query tracks.

5. EXPERIMENTAL RESULTS

The most obvious and straightforward way to represent a Web page is to merge body and title fields into a single field, which is used as the baseline in our study. In order to test the usefulness of keywords and description meta tags, we create three document representations for comparison: (*i*) body + title + description, (*ii*) body + title + keywords, and (*iii*) body + title + description + keywords.

Table 5 presents the effectiveness results of different document representations over the three TREC Web benchmark datasets. The top-ranked 100 documents are compared using the Normalized Discounted Cumulative Gain (NDCG) [53]. Statistical differences between different document representations are computed using the Student's t-test employed at 95% confidence level ($p < 0.05$).

[9] https://jsoup.org
[10] https://github.com/trec-web/trec-web-2014
[11] http://ir.cis.udel.edu/million/statAP_MQ_eval_v4.pl

Table 5. Comparison of different document representations over the GOV2, ClueWeb09B and ClueWeb12B collections and the associated query sets from TREC and NTCIR. Significant improvement or degradation with respect to the baseline document representation (body+title) is indicated († / «) (*p*-value < 0.05). Each figure shown in bold is the highest in that cell. Effectiveness is shown as NDCG at 100 documents returned (NDCG@100).

Model	Field	GOV2	MQ07	MQ08	MQ09	CW09B	CW12B	NTCIR
BM25	body+title	0.37337	**0.29864**	**0.32641**	0.32933	0.16719	**0.07652**	0.31750
	+description	**0.37384**	0.29845	0.32582	**0.33036**	**0.16784**	0.07627	**0.31766**
	+keywords	0.37167	0.29788	0.32281«	0.32603	0.16667	0.07501«	0.30981«
	+desc.+key.	0.37117	0.29751	0.32291«	0.32428	0.16513	0.07452	0.30801«
LM.DIR	body+title	0.39547	0.31852	0.32171	0.30092	0.18021	0.11466	0.33902
	+description	0.39564	0.31957†	**0.32209**	0.30735†	0.18213†	**0.11527**	**0.34393**†
	+keywords	0.39656	0.31997†	0.32097	0.30984†	0.18105	0.11394	0.33775
	+desc.+key.	**0.39709**	**0.32065**†	0.32118	**0.31228**†	**0.18237**	0.11385	0.34171
DFI	body+title	0.36204	0.26770	0.25407	0.30915	0.18648	0.09512	0.32798
	+description	0.36348†	0.26799	0.25460	0.31299†	**0.18740**	**0.09640**	**0.33190**
	+keywords	0.36334	0.26830	**0.25476**	0.31338†	0.18671	0.09376	0.32420
	+desc.+key.	**0.36380**	**0.26869**	0.25436	**0.31406**†	0.18562	0.09476	0.32569
LGD	body+title	0.37616	0.32032	0.34030	0.33102	0.17090	0.09672	0.33655
	+description	**0.37683**	**0.32044**	**0.34248**†	**0.33187**	0.17180	**0.09754**	**0.33821**
	+keywords	0.37418	0.31873«	0.33993	0.33048	**0.17237**	0.09562	0.32571«
	+desc.+key.	0.37460	0.31807«	0.33966	0.32991	0.17136	0.09532	0.32666«

The trends observed from Table 5 can be summarized as follows. Language Model with Dirichlet prior smoothing (LM.DIR) benefits the most from the inclusion of the description and keywords fields and the improvements are statistically significant most of the time. The retrieval effectiveness of the DFI model also increases with the inclusion of meta fields, but the improvements are not always statistically significant. The inclusion of keywords fields deteriorates the retrieval effectiveness of both BM25 and LGD models. Moreover, the observed decrease in effectiveness is statistically significant for half of the query sets.

The grand observation holding true for all query sets and term-weighting models is that the addition of description field never causes a significant decrease in retrieval effectiveness, and it increases the retrieval effectiveness most of the time. Furthermore, some of the improvements gained by the description field are statistically significant. Experimental results show that the inclusion of keywords field exhibits risk but it is safe to include description field.

This finding is aligned with the electronic articles[12] that discuss whether keywords and description fields are used by commercial search engines or not. It is claimed that although these fields are used to be utilized by search engines in the early years, keywords tag is abandoned nowadays due to *keyword stuffing*. Recall that the author of a page can insert *honeypot keywords* that would deceive commercial search engines and attract traffic. But, description tag is said to be still used to generate snippets on the search engine results page. It is worthwhile to note that this information is based on the experiences of Search Engine Optimization [54] community. Since the algorithms of commercial Web search engines are trade secrets and may change over time, this information is not certain.

6. DISCUSSION

The aim of this study is not to propose a comprehensive approach to field-based Web retrieval. But it is worthwhile to note that here are five approaches to combine different document representations in a retrieval system.

1) A meta-search [55] or data fusion [56] approach would treat each document representation as a distinct retrieval strategy, and then combine search results from the individual fields to produce a unified ranked list (i.e. post-retrieval). The motivation here is that finding all relevant documents for a given query is beyond the capability of a single retrieval strategy.

[12] https://support.google.com/webmasters/answer/79812

2) Term-weighting models are extended to obtain *field-based* models that can handle multiple weighted fields (e.g. BM25F [1] and PL2F [57]). These models combine the statistics of query terms obtained from different fields (i.e. pre-retrieval).

3) Recently, query terms' term-weighting scores calculated on different fields (e.g. BM25 of URL) are used as query-dependent features when learning a ranking model [16, 24, 25]. Figure 2 illustrates how different fields are utilized in a learning-to-rank dataset in such a way.

4) More recently, Zamani et al. [58] propose a neural ranking model that can handle multiple document representations. The authors introduce a field-level masking method to tackle the challenges of *missing fields* and *repeatable fields*, as well as a field-level dropout method to avoid relying too much on any one field.

5) A selective approach applied to different document representations on a per-query basis [59, 60]. Instead of trying to merge different result lists into a single ranked list, a selective approach tries to predict which single document representation alone would attain the highest effectiveness for a given query where each document representation is treated as a distinct retrieval strategy.

In this section, we present the effectiveness results of the individual fields and highlight the importance of a selective retrieval applied on different document representations. Selective IR is the study of predicting the most effective retrieval strategy on a per-query level amongst a set of existing alternatives, by contrast to the uniform application of a single strategy to all queries. When the effectiveness of retrieval strategies on individual queries is analyzed, it has been observed that different strategies attain highest scores at different queries (i.e. the best retrieval strategy varies across queries). A successful selective IR application is capable of attaining higher effectiveness than the individual strategies used alone without inventing a new one.

Table 6. Comparison of individual document representations over the ClueWeb09B dataset and the associated Million Query and Web Track query sets from TREC. Effectiveness is shown as NDCG at 100 documents returned (NDCG@100). The highest effectiveness score in each cell is shown in bold.

		MQ09		**CW09B**	
Model	**Field**	**NDCG**	**# Hits**	**NDCG**	**# Hits**
BM25	Oracle	**0.41985**	**562**	**0.22095**	**197**
	body	0.32417	290	0.16545	81
	title	0.18104	91	0.10543	29
	anchor	0.18961	71	0.13428	55
	URL	0.14769	42	0.08377	16
	keywords	0.15172	41	0.08111	11
	description	0.10015	27	0.06010	5
LM.DIR	Oracle	**0.36703**	**562**	**0.21815**	**197**
	body	0.29429	343	0.17625	111
	title	0.11732	66	0.07135	24
	anchor	0.14125	64	0.11052	40
	URL	0.13588	57	0.07153	18
	keywords	0.07505	16	0.04253	3
	description	0.06334	16	0.04070	1
DFI	Oracle	**0.40071**	**562**	**0.23174**	**197**
	body	0.30013	301	0.18218	101
	title	0.18709	96	0.11289	39
	anchor	0.16572	60	0.12192	29
	URL	0.14783	47	0.07996	16
	keywords	0.13960	32	0.07198	4
	description	0.10134	26	0.05961	8
LGD	Oracle	**0.41903**	**562**	**0.22257**	**197**
	body	0.32308	307	0.16806	83
	title	0.19222	87	0.11644	36
	anchor	0.18979	69	0.13483	52
	URL	0.15200	42	0.08345	15
	keywords	0.14912	32	0.07917	8
	description	0.10579	25	0.06225	3

Table 6 lists the effectiveness scores of individual fields as well as an oracle method that always predicts the best field on a per-query basis. The intent is to emphasize the substantial improvement in retrieval effectiveness, at least in theory, that could be achieved by a selective approach that predicts what field should be applied to which query. The '#Hits' column in Table 6 shows the number of times a document representation attains the highest effectiveness score (i.e. ranks first).

The most effective single field is the body field whereas description and keywords fields are two of the worst effective fields in Table 6. Although both description and keywords fields perform poorly when used individually to serve all queries, keywords field attains the best effectiveness for 5% of the queries. This finding suggests that keywords field is not totally useless. Furthermore, description field is better than keywords when appended to title and body fields, while, on the contrary, keywords field is better than description when used individually.

As shown in Table 6, a selective oracle method is in theory capable of bringing nearly 29% improvement over the most effective single field (i.e. body), on average. Here, keywords or description fields attain the best effectiveness for 8.5% of the queries. Indeed, the utilization of keywords and description fields would contribute to the effectiveness of an oracle method that always predicts the best field with 100% accuracy on a per-query basis.

7. CONCLUSION

While different document representations (URL, title, body, anchor) are mentioned in the Web retrieval literature, there has been little empirical investigation into whether meta tags are useful for improving retrieval effectiveness or not. In this paper, we investigate the benefit of keywords and description meta tags, when combined with body plus title using the state-of-the-art term-weighting models.
Our results show that the addition of description field could significantly improve the retrieval effectiveness, while, in contrast, the addition of keywords field could significantly deteriorate the retrieval effectiveness (i.e. exhibits significant risk).

The present study leverages meta elements in a static retrieval setting (i.e. no machine learning is employed). Our future work will experiment with the other approaches explained in the previous section, in which meta elements can be leveraged. For example, in a learning-to-rank setting, a new set of features can be constructed from term-weighting models' scores calculated on keywords and description fields (e.g. BM25 of keywords) and these new features can be used to extend the standard set of query-dependent features (shown in Fig. 2) that are based on URL, title, body, and anchor.

REFERENCES

[1] Robertson S, Zaragoza H, Taylor M. Simple BM25 Extension to Multiple Weighted Fields, in Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, pp. 42-49.

[2] Croft WB. "Combining Approaches to Information Retrieval," W. B. Croft, Ed., ed: Springer US, 2000, pp. 1-36.

[3] Turner TP, Brackbill L. Rising to the top: evaluating the use of the HTML meta tag to improve retrieval of World Wide Web documents through Internet search engines. Library Resources & Technical Services 1998; 42: 258-271.

[4] Hiemstra D, Hauff C, "MapReduce for Information Retrieval Evaluation: "Let's Quickly Test This on 12 TB of Data"," in *Multilingual and Multimodal Information Access Evaluation*, M. Agosti,

N. Ferro, C. Peters, M. de Rijke, and A. Smeaton, Eds., ed: Springer Berlin Heidelberg, 2010, pp. 64-69.

[5] Mao J, Sakai T, Luo C, Xiao P, Liu Y, Dou Z. Overview of the NTCIR-14 we want web task. 2019; 455-467.

[6] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 1998; 30: 107-117.

[7] Ounis I, Amati G, Plachouras V, He B, Macdonald C, Johnson D. Terrier Information Retrieval Platform, in Advances in Information Retrieval, pp. 517-519.

[8] Yang P, Fang H, Lin J. Anserini: Reproducible Ranking Baselines Using Lucene. J. Data and Information Quality 2018; 10: 16:1-16:20.

[9] Verma M, Yilmaz E, Mehrotra R, Kanoulas E, Carterette B, Craswell N, et al. Overview of the TREC Tasks Track 2016. 2016.

[10] Sanderson M, Croft WB. The History of Information Retrieval Research. Proceedings of the IEEE 2012; 100: 1444-1451.

[11] Craswell N, Hawking D, Robertson S. Effective Site Finding Using Link Anchor Information, in Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; New Orleans, Louisiana, USA; 2001, pp. 250-257.

[12] Eiron N, McCurley KS, "Analysis of anchor text for web search," presented at the Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, Toronto, Canada, 2003.

[13] Kraft R, Zien J, "Mining anchor text for query refinement," presented at the Proceedings of the 13th international conference on World Wide Web, New York, NY, USA, 2004.

[14] Dang V, Croft BW, "Query reformulation using anchor text," presented at the Proceedings of the third ACM international conference on Web search and data mining, New York, New York, USA, 2010.

[15] Anh VN, Moffat A. The Role of Anchor Text in ClueWeb09 Retrieval. 2010.

[16] Macdonald C, Santos RLT, Ounis I. The whens and hows of learning to rank for web search. Information Retrieval 2013; 16: 584-628.

[17] Kang I-H, Kim G. Query Type Classification for Web Document Retrieval, in Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 64-71.

[18] Song R, Wen J-R, Shi S, Xin G, Liu T-Y, Qin T, et al. Microsoft Research Asia at Web Track and Terabyte Track of TREC 2004. 2004.

[19] Ogilvie P, Callan J. Combining Structural Information and the Use of Priors in Mixed Named-Page and Homepage Finding. 2003.

[20] Westerveld T, Kraaij W, Hiemstra D. Retrieving web pages using content, links, urls and anchors. 2001.

[21] Chibane I, Doan B-L. A Web Page Topic Segmentation Algorithm Based on Visual Criteria and Content Layout, in Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 817-818.

[22] Craswell N, Hawking D. Overview of the TREC-2004 Web Track. 2004.

[23] Zheng G, Callan J, "Learning to Reweight Terms with Distributed Representations," presented at the Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, 2015.

[24] Qin T, Liu T-Y, Xu J, Li H. LETOR: A benchmark collection for research on learning to rank for information retrieval. Information Retrieval 2010; 13: 346-374.

[25] Macdonald C, Santos RLT, Ounis I, He B. About Learning Models with Multiple Query-dependent Features. ACM Trans. Inf. Syst. 2013; 31: 11:1-11:39.

[26] Collins-Thompson K, Ogilvie P, Zhang Y, Callan J. Information filtering, novelty detection, and named-page finding. 2002.

[27] Ogilvie P, Callan J, Callan J. Combining Document Representations for Known-item Search, in Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 143-150.

[28] Savoy J, Rasolofo Y. Report on the TREC 11 experiment: Arabic, named page and topic distillation searches. 2002.

[29] Zhou Z, Guo Y, Wang B, Cheng X, Xu H, Zhang G. TREC 2004 Web Track Experiments at CAS-ICT. 2004.

[30] Tomlinson S. Robust, web and terabyte retrieval with hummingbird searchserver at TREC 2004. 2004.

[31] Wen J-R, Song R, Cai D, Zhu K, Yu S, Ye S, et al. Microsoft Research Asia at the Web Track of TREC 2003. 2003.

[32] Roy D, Mitra M, Ganguly D. To Clean or Not to Clean: Document Preprocessing and Reproducibility. J. Data and Information Quality 2018; 10: 18:1-18:25.

[33] Gadge J, Bhirud S. Contextual weighting approach to compute term weight in layered vector space model. Journal of Information Science 2019; DOI: 10.1177/0165551519860043.

[34] Spirin N, Han J. Survey on web spam detection: principles and algorithms. SIGKDD Explor. Newsl. 2012; 13: 50-64.

[35] Lewandowski D. Web searching, search engines and Information Retrieval. Inf. Serv. Use 2005; 25: 137-147.

[36] Craven TC. Variations in use of meta tag descriptions by Web pages in different languages. Information Processing & Management 2004; 40: 479-493.

[37] Craven TC. Variations in Use of Meta Tag Keywords by Web Pages in Different Languages. Journal of Information Science 2004; 30: 268-279.

[38] Zhang J, Jastram I. A study of metadata element co-occurrence. Online Information Review 2006; 30: 428-453.

[39] Alimohammadi D. Meta-tags: still a matter of opinion. The Electronic Library 2005; 23: 625-631.

[40] Clarke C, Craswell N, Soboroff I. Overview of the TREC 2004 Terabyte Track. 2004.

[41] Callan J, Hoy M, Yoo C, Zhao L. (2009, The ClueWeb09 Dataset. Available: http://boston.lti.cs.cmu.edu/classes/11-742/S10-TREC/TREC-Nov19-09.pdf

[42] Callan J. (2012, The Lemur Project And its ClueWeb12 Dataset. Available: http://opensearchlab.otago.ac.nz/SIGIR12-OSIR-callan.pdf

[43] Luo C, Sakai T, Liu Y, Dou Z, Xiong C, Xu J. Overview of the NTCIR-13 we want web task. 2017; 394-401.

[44] Robertson S, Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends(®) in Information Retrieval 2009; 3: 333-389.

[45] Zhai C, Lafferty J. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. ACM Trans. Inf. Syst. 2004; 22: 179-214.

[46] Kocabaş İ, Dinçer BT, Karaoğlan B. A nonparametric term weighting method for information retrieval based on measuring the divergence from independence. Information Retrieval 2014; 17: 153-176.

[47] Clinchant S, Gaussier É. Information-based Models for Ad Hoc IR, in Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 234-241.

[48] Białecki A, Muir R, Ingersoll G. Apache Lucene 4, in Proceedings of the SIGIR 2012 Workshop on Open Source Information Retrieval, pp. 17-24.

[49] Azzopardi L, Crane M, Fang H, Ingersoll G, Lin J, Moshfeghi Y, et al. The Lucene for Information Access and Retrieval Research (LIARR) Workshop at SIGIR 2017, in Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1429-1430.

[50] McCandless M, Hatcher E, Gospodnetic O. Lucene in Action, Second Edition: Covers Apache Lucene 3.0, Manning Publications Co., 2010.

[51] Krovetz R. Viewing Morphology As an Inference Process, in Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 191-202.

[52] Carterette B, Pavlu V, Fang H, Kanoulas E. Million Query Track 2009 Overview. 2009.

[53] Järvelin K, Kekäläinen J. Cumulated Gain-based Evaluation of IR Techniques. ACM Trans. Inf. Syst. 2002; 20: 422-446.

[54] Khan MNA, Mahmood A. A distinctive approach to obtain higher page rank through search engine optimization. Sādhanā 2018; 43: p. 43.

[55] Aslam JA, Montague M. Models for Metasearch, in Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 276-284.

[56] Montague M, Aslam JA. Condorcet Fusion for Improved Retrieval, in Proceedings of the Eleventh International Conference on Information and Knowledge Management, pp. 538-548.

[57] Macdonald C, Plachouras V, He B, Lioma C, Ounis I, "University of Glasgow at WebCLEF 2005: Experiments in Per-Field Normalisation and Language Specific Stemming," in *Accessing Multilingual Information Repositories*, C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, *et al.*, Eds., ed: Springer Berlin Heidelberg, 2006, pp. 898-907.

[58] Zamani H, Mitra B, Song X, Craswell N, Tiwary S. Neural Ranking Models with Multiple Document Fields, in Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining; Los Angeles, California, USA; 2018, pp. 700-708.

[59] Plachouras V, Ounis I, Cacheda F. Selective Combination of Evidence for Topic Distillation Using Document and Aggregate-level Information, in Proceedings of the RIAO 2004 - Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval, pp. 610-622.

[60] Plachouras V, Cacheda F, Ounis I. A decision mechanism for the selective combination of evidence in topic distillation. Information Retrieval 2006; 9: 139-163.