information

Article

Improving Consumer Health Search with Field-Level Learning-to-Rank Techniques

Hua Yang [1,*] and Teresa Gonçalves [2,3]

[1] School of Artificial Intelligence, Zhongyuan University of Technology, Zhengzhou 450007, China
[2] Department of Computer Science, University of Évora, 7000-671 Évora, Portugal; tcg@uevora.pt
[3] VISTA Lab, Algoritmi Center, University of Évora, 7000-671 Évora, Portugal
* Correspondence: huayang@zut.edu.cn

Abstract: In the area of consumer health search (CHS), there is an increasing concern about returning topically relevant and understandable health information to the user. Besides being used to rank topically relevant documents, Learning to Rank (LTR) has also been used to promote understandability ranking. Traditionally, features coming from different document fields are joined together, limiting the performance of standard LTR, since field information plays an important role in promoting understandability ranking. In this paper, a novel field-level Learning-to-Rank (f-LTR) approach is proposed, and its application in CHS is investigated by developing thorough experiments on CLEF' 2016-2018 eHealth IR data collections. An in-depth analysis of the effects of using f-LTR is provided, with experimental results suggesting that in LTR, title features are more effective than other field features in promoting understandability ranking. Moreover, the fused f-LTR model is compared to existing work, confirming the effectiveness of the methodology.

Keywords: health informatics; consumer health search; information retrieval; learning to rank; understandability

Citation: Yang, H.; Gonçalves, T. Improving Consumer Health Search with Field-Level Learning-to-Rank Techniques. *Information* **2024**, *15*, 695. https://doi.org/10.3390/info15110695

Academic Editor: Neil Vaughan

Received: 19 September 2024
Revised: 28 October 2024
Accepted: 29 October 2024
Published: 3 November 2024

1. Introduction

Consumer health search (CHS) [1–3], also known as consumer health information retrieval (CHIR), is one research area in information retrieval (IR) that aims to search health information specifically for non-expert users. For example, in the context of consumer health search, a layperson experiencing symptoms such as a cough and fever might enter a query like "*What causes cough and fever?*". In contrast, an expert health search would involve more technical terminology, with a query articulated as "*Differential diagnosis for suspected respiratory infections*".

One notable challenge in the CHS area is that non-expert consumers have difficulty understanding the retrieved answers. A topically relevant document may not help a consumer if the document is beyond his/her understandability level [2,4]. In health-related areas, the readability of written texts regarding appointments, medication, and medication doses is essential for a reader; poor understandability of these texts is associated with poor health outcomes and may include increased mortality [4]. If the user finds the retrieved document difficult to understand, even if it is highly relevant, they are likely to give up and move on to another one [2]. This important need constitutes the motivation of this work: retrieving not only topically relevant but also understandable results for consumers in the CHS task.

In this work, Learning-to-Rank (LTR) techniques [5] are studied to improve understandability beyond topicality relevance in the area of CHS. The main contributions of our work can be featured as follows: (i) We propose the field-level Learning-to-Rank (f-LTR) model; different from the standard LTR approach where one single model is trained, in f-LTR, a set of Learning-to-Rank models are learned, with each one emphasizing information taken from one specific field; (ii) We evaluate and prove its effectiveness in surpassing

the state-of-the-art techniques; thorough experimental tests of f-LTR models are conducted leading to improve understandability ranking in the area of CHS.

The rest of the paper is organized as follows: Section 2 reviews the related work in Learning to Rank and understandability in CHS; Section 3 details the proposed method; Section 4 describes the conducted experiments and Section 5 analyzes the results; finally, Section 6 concludes and presents suggestions for future work.

2. Literature Review

The f-LTR approach proposed in this paper builds upon previous work developed regarding understandability research in CHIR and feature-based LTR research.

2.1. Understandability in Consumer Health Search

The research community has shown significant interest in understanding online health information [6]. Health information retrieval concerns different areas, from biomedical literature retrieval for clinical cases to health-related retrieval by general non-expert users. Many existing IR systems merely consider the topical relevance of the retrieved documents without taking into account the dimension of understandability. A topically relevant but not understandable document is of no value to a consumer. In the health domain search, this is even more important, since non-understandable information may cause other issues. To increase the access and utility of health-related information to the public, organizations recommend a specific readability level for health information.

The American Medical Association (AMA) recommends a sixth-grade reading level, and the United States National Institute of Health (NIH) recommends that print materials for the public should use plain language with a target readability equivalent to the sixth-grade level and no greater than eighth-grade [7,8]. A study analyzing the results from 70 websites on a popular search engine for the health query "*congestive heart failure*" found that only 7.1% of the documents met the recommended sixth-grade reading level based on one assessment tool. Moreover, none of the websites achieved a sixth-grade reading level when evaluated using all five assessment tools [9]. Another work found that no article abstract met the NIH readability target of sixth grade or below, and only one was below the recommended ceiling of eighth-grade equivalent [10].

Computational readability assessments have been developed to automatically evaluate the reading level of a given text; for example, Simple Measure of Gobbledygook (SMOG) was the preferred measure of readability when evaluating consumer-oriented health care material [11,12]. Other popular computational readability assessments for web health documents include the Flesch reading ease, Flesch–Kincaid grade level, Gunning Fog index, and Coleman–Liau index [9]. These measures are based on the surface characteristics of a document, such as sentence length and word length of syllables.

2.2. Learning-to-Rank Techniques

In the IR area, Machine Learning techniques can be applied to build ranking models for the information retrieval systems, and this is known as Learning to Rank [13–15]. Training queries, related documents, and the matching relevance judgments for the query and document pairs typically comprise the training data. The learning algorithms are then used to generate an LTR model. Similarly, the creation of testing data for evaluation, which includes test queries and associated documentation, follows a methodology analogous to that used in the generation of training data. IR and LTR models collaborate to sort documents based on their relevance as answers to questions, thereby generating a ranked list of documents that respond to the query.

LTR approaches have been studied in many health search contexts such as expert medical search by physicians, Electronic Health Records search by patients, and consumer health search by laypeople [16–18]. One focal research in Learning to Rank is exploring valuable features; depending on the application, different features can be extracted and used in training a Learning-to-Rank model [19,20].

Traditionally, potentially effective features are extracted and naively combined together to create a feature list, with most studies exploring new features but joining all into a single list [21–23]. Little attention is given to how these features should be grouped, such as constructing multiple feature lists instead of relying on a single, consolidated list.

3. Methodology

In this section, the hypothesis of different fields contributing differently is introduced, followed by a detailed explanation of the proposed f-LTR approach. Finally, the rationale behind the chosen methodology is discussed.

3.1. Hypothesis

Document fields are assumed to contribute differently to improving the effectiveness of information retrieval. As an example, Table 1 presents fields that represent standard sections (such as the heading, title, and body) of an HTML web document.

Table 1. Typical fields of an HTML document.

Field	Description
H1–H6	Section headings at different levels; H1 is the highest-level heading and H6 is the lowest level.
Title	A document title.
Header	Defines a header for a document or section.
Meta	Metadata of a document such as author, publication date, keywords, etc.
Anchor	Anchors a URL to some text on a web page.
Body	Body content of a document.
Else	Not defined in any field.
Whole	The contents of the full document.

The hypotheses are as follows: (i) In training LTR models, the naive combination of features, which joins features extracted from different fields into a single feature list, may decrease the contribution of the field information. Training LTR models using grouped, field-level features is expected to be more efficient. (ii) The fusion of results from a set of pre-trained field-specific LTR models is anticipated to be more effective than a model trained with features that are naively combined from multiple fields. The f-LTR approach is proposed to validate these assumptions.

3.2. F-LTR Approach

The architecture of the proposed f-LTR applied in the CHS task is presented in Figure 1.

Standard Learning to Rank works at the document level, combining all extracted features from various fields into a single list to train one LTR model. In contrast, the proposed f-LTR model operates at the field level, enabling a more refined approach that distinctively highlights and prioritizes features specific to individual fields. Once the f-LTR model is trained, it can be employed to rank results for new queries.

However, a notable limitation of the LTR model is its reliance on a single field during the training process, which may result in biased outcomes.

To address this issue, the f-LTR approach is proposed and illustrated in Figure 1. The f-LTR approach, aligned with the methodologies of state-of-the-art information retrieval (IR) techniques, mainly includes two stages. In the first stage, the features are grouped by specific fields, and a set of f-LTR models is created, with each model trained using the features of a specific field. In the second stage, the scores generated by the pre-trained f-LTR models are fused using a designated fusion method.

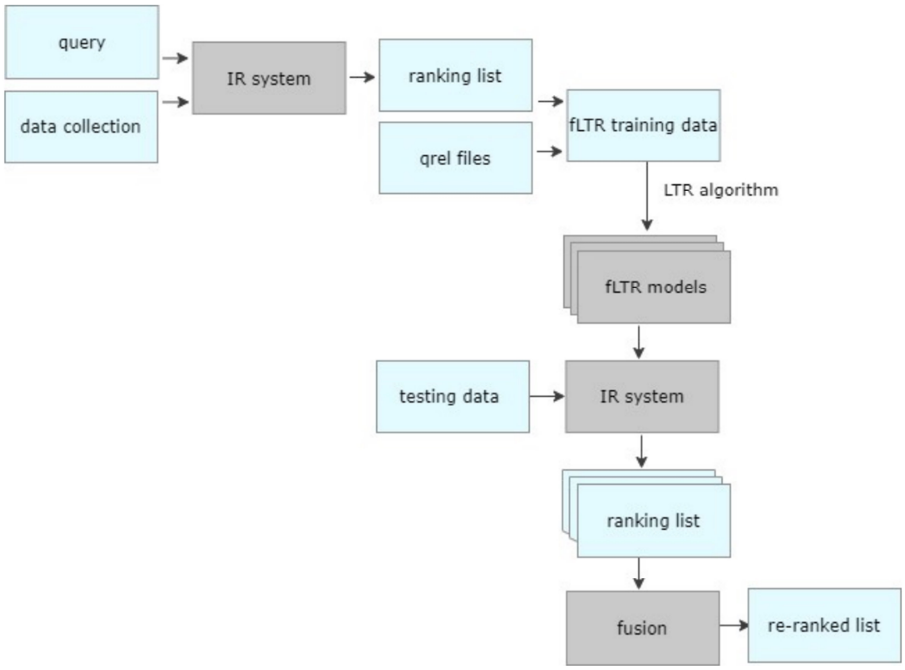

Figure 1. The architecture of the proposed f-LTR approach in CHS.

3.2.1. Stage One: Building f-LTR Models

In the first stage, features are grouped by specific fields, resulting in the creation of a set of f-LTR models. Each model is trained using a distinct group of features extracted from a single field. The framework for this first stage is illustrated in Figure 2.

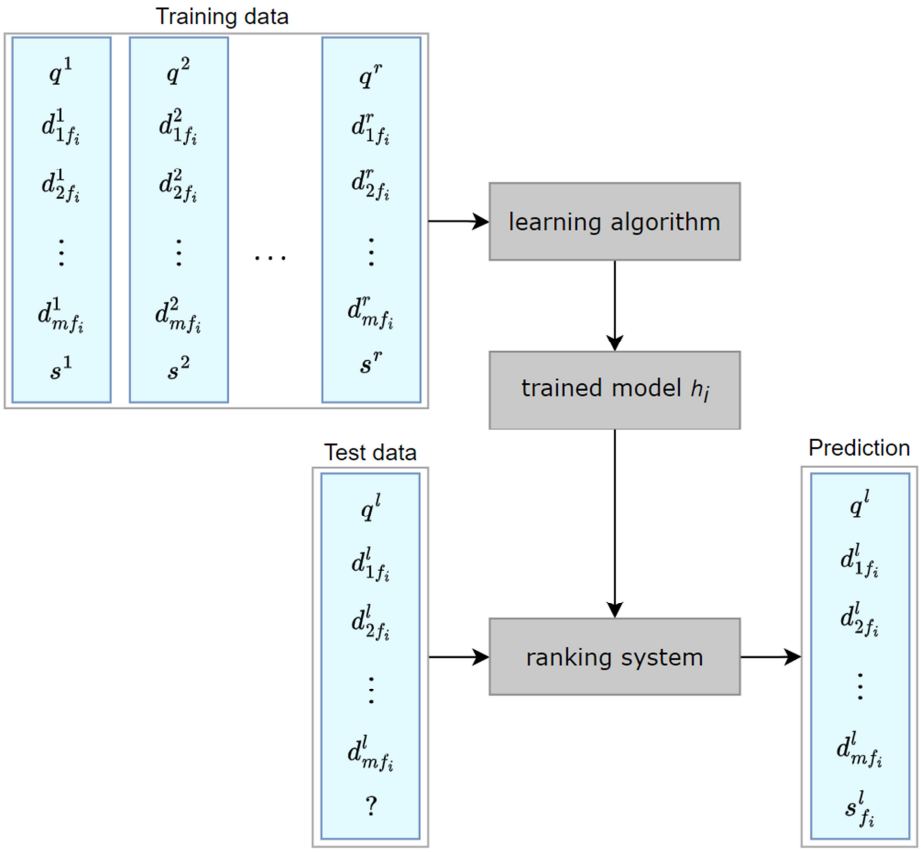

Figure 2. The f-LTR framework.

As shown in Figure 2, the training data consist of r queries $q^j (j = 1, \ldots, r)$. A number of documents ($d^j_{1f_i}$, $d^j_{2f_i}$, …, $d^j_{mf_i}$) are associated to each query q^j; f_i indicates that the features are extracted from the ith field of a document. The training data also include the corresponding relevance judgments s^j for each query and document pair. Employing the learning algorithm, an f-LTR model h_i is learned from the training data. In the testing phase, given a new query q^l, the trained model h_i joined with the ranking system is used to rank the documents. A ranking list of relevant documents $s^l_{f_i}$ is produced for the query q^l.

Building on the aforementioned concepts, let us consider that features are extracted from n fields: $f_1, f_2, \ldots, f_n$. For each field, a corresponding Learning-to-Rank model h is constructed. This results in the creation of n models $(h_1, h_2, \ldots, h_n)$. Because these models are developed using features from a specific field, they are referred to as f-LTR models.

3.2.2. Stage Two: Fusion of f-LTR Models

Following the approach outlined in the first stage, a group of n f-LTR models is trained, with each model learning from a single field. Consequently, n corresponding ranking lists are generated, denoted as $s_{f_1}, s_{f_2}, \ldots, s_{f_n}$. In the subsequent fusion stage, these ranking lists are combined using a designated fusion method, resulting in a final re-ranked list presented to the user. The second stage, also referred to as the fusion stage, is illustrated in Figure 3.

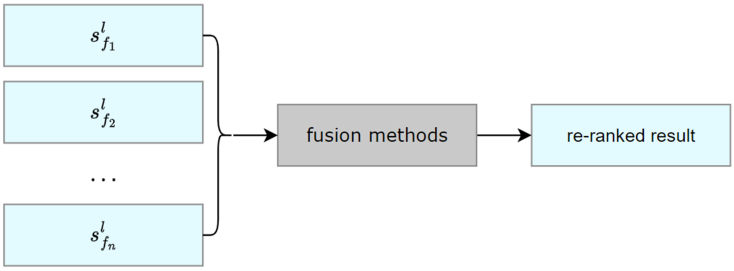

Figure 3. Fusion of f-LTR models.

3.3. Rationale for the Methodology

The objective is to rank documents within the corpus D for a given query q. Assuming a document d, (where $d \in D$) consists of n fields, each field is denoted as f_i $(i = 1, 2, \ldots, n)$. The f-LTR model trained using information from the field f_i is represented as h_i. The score of the document d in relation to the query q as assessed by the f-LTR model h_i is denoted as:

$$s_{f_i}(d, q, h_i)(i = 1, 2, \ldots, n) \tag{1}$$

Since the result s_{f_i} is derived from a single field, a total of n f-LTR models are trained, with each model utilizing features extracted from one specific field. The fusion algorithms are then applied to the results obtained from this set of f-LTR models. The fusion algorithm is defined as the function $G[x]$, and the fused score S is expressed as:

$$S = G[\, s_{f_1}(d, q, h_1), \ldots,\, s_{f_i}(d, q, h_i), \ldots, s_{f_n}(d, q, h_n)] \tag{2}$$

Let the weight assigned to each field be denoted as w_i. The fused score S of the document d can be expressed as:

$$S = G[\, w_1 s_{f_1}(d, q, h_1), \ldots,\, w_i s_{f_i}(d, q, h_i), \ldots, w_n s_{f_n}(d, q, h_n)] \tag{3}$$

With $G[x]$ as a linear function, the fused score is denoted as:

$$S = \sum_{i=1}^{n} w_i s_{f_i}(d, q, h_i) \tag{4}$$

If all field weights are equal, the fused score S of the document d can be expressed as:

$$S = \sum_{i=1}^{n} s_{f_i}(d, q, h_i) \tag{5}$$

This formula represents the average of the scores obtained from each field's f-LTR model and is known as the CombSUM calculation (combined sum). It is a simple and widely used score fusion technique. In the context of IR, CombSUM works by aggregating the individual scores from different sources to produce a final combined score and assumes that all sources contribute equally to the final score.

In addition to CombSUM, other effective fusion methods include CombMax, CombMin, CombANZ, CombMNX, and CombMed [24,25]. Fusion through a linear combination of scores has been shown to be efficient and is widely used in various score-based fusion tasks within IR [26–28]. Given its effectiveness, we adopted them as the fusion techniques in this work. The formulas are illustrated in Equations (6)–(10):

$$CombMax = \max(s_{f_1}(d, q, h_1), s_{f_2}(d, q, h_2), \ldots, s_{f_n}(d, q, h_n)) \tag{6}$$

$$CombMin = \min(s_{f_1}(d, q, h_1), s_{f_2}(d, q, h_2), \ldots, s_{f_n}(d, q, h_n)) \tag{7}$$

$$CombANZ = \frac{\sum_{i=1}^{n} s_{f_i}(d, q, h_i)}{k} \tag{8}$$

$$CombMNX = \min(s_{f_i}(d, q, h_i) \mid s_{f_i}(d, q, h_i) > 0) \tag{9}$$

$$CombMed = median(s_{f_1}(d, q, h_1), s_{f_2}(d, q, h_2), \ldots, s_{f_n}(d, q, h_n)) \tag{10}$$

4. Experiments

In this section, the dataset collections used in the study are introduced first. Then, the features and the ranker for the experiments are described. Lastly, the evaluation metrics and tools employed for performance assessment are presented.

4.1. Datasets

The ideal data collection for carrying out experimental work should include queries generated by non-expert consumers, retrieved documents focused on health or medical topics (excluding scientific biomedical literature), high-quality assessments of query–document pairs conducted by medical experts, and large datasets well suited for IR tasks. Nonetheless, acquiring high-quality data that meet these standards is not a straightforward task, mainly because of the following: (i) Not too many open dataset collections are available since it is a specific topic in the health search area where data privacy needs to be assured; (ii) The assessment of the query–document pairs is costly since they require evaluators to have strong medical knowledge backgrounds (which usually only experts or majors in the medical area have).

To meet the needs of our research goal, datasets that included both topical relevance and understandability assessment results were chosen. In total, two dataset collections were used: CLEF' 2016–2017 and CLEF'2018 eHealth IR datasets. CLEF eHealth is an evaluation lab that has organized evaluation campaigns in the medical and biomedical domain since 2013. The CLEF eHealth task follows the TREC-style evaluation process and provides a shared and standard IR data collection that contains a dataset, a query set, and assessment files. Complete data collections and evaluation frameworks were available, so the proposed approach was successfully tested on these data collections.

The details of the two data collections are summarized in Table 2 and include the dataset used for retrieving documents, the set of queries, and the associated Qrels. A Qrel file, short for query relevance file, specifies the relevance of documents to particular queries.

It is commonly used in the evaluation of IR systems to assess how effectively the system retrieves relevant documents in response to a given set of queries.

Table 2. Statistics of the data collections for our experiments.

	CLEF' 2016–2017 Collection	CLEF' 2018 Collection
Dataset	ClueWeb12-B13	5,535,120 Web pages
Query set	300	50
Qrels files	269,232	18,763

CLEF' 2016 and CLEF' 2017 collections included the same dataset and query set, but CLEF' 2017 has an increased assessment pool with more query–documents pairs. The two assessment files were combined into one for this work and named CLEF' 2016–2017 Collection. The query set was issued by the public (https://www.reddit.com/r/AskDocs, accessed on 28 October 2024) and expressed their real health information needs. A total of 300 medical questions resembling lay people's health queries were produced after preferred posts were chosen to serve as basic queries. CLEF' 2016 and CLEF' 2017 collections used ClueWeb12-B13 (https://huggingface.co/datasets/irds/clueweb12_b13_clef-ehealth, accessed on 28 October 2024) as the dataset, which contained about 52 million web pages. The query–document pairs were assessed by senior medical students. Relevance between a query and a document was graded as *highly relevant*, *somewhat relevant*, and *not relevant*. The collected understandability assessments ranged from 0 to 100, with 0 being the hardest to understand and 100 the easiest.

The CLEF' 2018 collection was generated following the same procedure as the CLEF' 2016 and CLEF' 2017 collections. It contains 5,535,120 web pages obtained from the CommonCrawl (http://commoncrawl.org, accessed on 28 October 2024) as the dataset, 50 medical queries issued by the public and gathered from the Health on the Net search engine, and 18,763 query–documents pairs for the assessment.

4.2. Features Extracted

Inspired by the work of Ru et al. [29] and Bhagawati and Subramanian [30], four fields were considered, namely, `Title`, `H1`, `Else` (the latter represents the texts that do not belong to `Title` or `H1`) and the full text of the document (dhe full text is regarded as one field information as well). Nine features that performed well in previous consumer health search studies [31,32] were taken into account; these nine extracted features were mostly based on classic IR models: `TFIDF` along with `TF` and `IDF`, the probabilistic model `BM25`, the language models `HiemstraLM` and `DirichletLM`, `BB2`, `PL2`, and `Dl`. A total of 36 features were then extracted (four fields with nine features each). An overview of the extracted features is presented in Table 3.

Table 3. Labels of the features extracted and grouped for f-LTR processing.

Feature	Field Group			
	Title (T)	H1 (H)	Else (E)	Full Doc (F)
TFIDF	T-TFIDF	H-TFIDF	E-TFIDF	F-TFIDF
TF	T-TF	H-TF	E-TF	F-TF
IDF	T-IDF	H-IDF	E-IDF	F-IDF
BM25	T-BM25	H-BM25	E-BM25	F-BM25
HiemstraLM	T-HiemstraLM	H-HiemstraLM	E-HiemstraLM	F-HiemstraLM
DirichletLM	T-DirichletLM	H-DirichletLM	E-DirichletLM	F-DirichletLM
BB2	T-BB2	H-BB2	E-BB2	F-BB2
PL2	T-PL2	H-PL2	E-PL2	F-PL2
Dl	T-D1	H-D1	E-D1	f-D1f

4.3. Developed Rankers

A group of experiments was designed and carried out on the two data collections. During the first stage, four f-LTR rankers were trained using the CLEF' 2016–2017 collection: R_T, R_H, R_E, and R_F, each trained using features from a specific field as presented in Section 4.2; R_A was built following the standard LTR approach and using all 36 features. During the second stage, these f-LTR rankers were used to perform retrieval on the CLEF' 2018 collection, and the scores were aggregated following six different strategies, namely, CombMAX, CombMIN, CombSUM, CombANZ, CombMNZ, and CombMED [33]. Three groups of aggregated rankers were generated, namely, R_{TH}, R_{THE}, and R_{THEF}, each using the six strategies, totaling 18 aggregated rankers, as described in Table 4.

Table 4. Description of developed rankers.

Ranker	Method Description
R_T	f-LTR model with F_{Title} features
R_H	f-LTR model with F_{H1} features
R_E	f-LTR model with F_{Else} features
R_F	f-LTR model with F_{full} features
R_A	f-LTR model with 36 features
R_{TH}	Fusion of R_T and R_H
R_{THE}	Fusion of R_T, R_H, and R_E
R_{THEF}	Fusion of R_T, R_H, R_E, and R_F

The PyTerrier retrieval platform [34] served as the primary environment for conducting the experiments, utilizing its LTR framework. All queries were pre-processed by converting characters to lowercase, removing stop words, and applying stemming with the Porter Stemmer.

The Okapi BM25 retrieval model was used to build the ranking models and all the parameters were set to default values (b = 0.75, k1 = 1.2, and k3 = 8), as recommended by Aloteibi [35]. When training an f-LTR model, up to 1000 documents per query were retrieved during the retrieval process. All models were trained and tuned with separate training and validation sets from the CLEF' 2016–2017 data collection and tested on CLEF' 2018 data.

4.4. Evaluation Metrics

The developed rankers were evaluated in terms of topical relevance as well as understandability. In terms of topical relevance, the three most important and frequently used assessment measures in information retrieval were included, namely, P@10 (Precision at 10), NDCG@10 (Normalized Discounted Cumulative Gain at 10), and MAP (Mean Average Precision). The rankers were evaluated at position 10 since users of online search engines are more likely to pay attention to the first 10 of the retrieved results. In assessing understandability relevance, two measures were included: uRBP and uRBPgr; uRBP uses binary understandability assessments, and uRBPgr uses graded understandability assessments [36].

Common assessment tools were utilized to calculate the aforementioned measures. The standard TREC competitions tool trec_eval (https://trec.nist.gov/trec_eval, accessed on 28 October 2024), was used for evaluating topical relevance; the Ubire tool (https://github.com/ielab/ubire, accessed on 28 October 2024), an understandability-biased IR evaluation tool, was employed for the understandability assessment.

5. Results

In this section, results on how field-based LTR models were able to equal or surpass the standard LTR model with much fewer features (detailed in Section 5.1) are analyzed,

and then the performance of the fused f-LTR rankers using different combinations and fusion methods is presented.

5.1. Comparing f-LTR Model to the Standard LTR Model

The four f-LTR rankers R_T, R_H, R_E, R_F, and the standard LTR ranker R_A were evaluated and the results are shown in Table 5. All generated ranking lists were evaluated using the assessment files with the introduced evaluation metrics.

Table 5. Comparison between f-LTR and standard LTR rankers.

Algorithm	Ranker	Understandability		Topicality		
		uRBP	**uRBPgr**	**P@10**	**NDCG@10**	**MAP**
f-LTR	R_T	**0.7131**	**0.3020**	**0.6820**	**0.6131**	**0.2428**
	R_H	0.6493	0.2630	0.6340	0.5683	0.2279
	R_E	0.5849	0.2380	0.5700	0.4753	0.2115
	R_F	0.6539	0.2750	0.6620	0.5395	0.2404
Standard LTR	R_A	0.5821	0.2550	0.6420	0.5687	0.2177

As can be observed, the best performance was achieved by the f-LTR ranker R_T. R_T surpassed the standard ranker R_A with improvements of 22.5% in uRBP and 18.4% in uRBPgr. Similar results were observed with the topical relevance evaluation metrics: R_T surpassed R_A with improvements of 6.2% in P@10, 7.8% in NDCG@10, and 11.5% in MAP.

Turning our attention to the other f-LTR rankers, R_H and R_F were also able to surpass R_A in most assessment metrics; only R_E was not able to exceed it.

The result suggests that using features selected from one field can achieve similar and even better performance than using features from the full document. In the experiments, the f-LTR model employed one-quarter of the features when compared to the standard LTR model.

Comparing the performance among the four f-LTR rankers (R_T, R_H, R_E, and R_F), the title-based ranker, R_T, was the most effective one. This suggests that not all fields contribute the same when building an LTR model. Features extracted from the *Title* field proved to be the most effective ones, followed by *H1* and *full* field information, with the *Else* field being the worst. This suggests that selecting specific features can lead to the development of more effective LTR models.

5.2. Fused f-LTR Rankers

The experiments conducted with the f-LTR model offered valuable insights into the effectiveness of features derived from field information, allowing for the identification of the most impactful field for training the f-LTR model.

However, these findings were not adequate to fully assess the performance of the comprehensive retrieval model. A single f-LTR ranker does not incorporate information from other fields, which is essential for the effective design of a retrieval model. To quantify this difference, the title-based ranker R_T was employed as the primary ranker and was gradually combined with other f-LTR rankers. R_T was selected because it demonstrated the best performance among the four f-LTR rankers.

Meanwhile, six different fusion methods (see Section 4.3) were tested for each ranker combination, generating a total of 18 combined models. These combined rankers were assessed, and the best-performing ones are presented in Table 6. CombMED and CombSUM performed the same and achieved the best scores when compared to the other four fusion methods. CombMED was used as the representative and is denoted as *med*.

An increase in retrieval performance was observed when gradually fusing results of R_T with other f-LTR rankers. To further investigate this phenomenon, an empirical evaluation was conducted. In this paper, the *uRBP* metric was used as an example; however, similar

behaviors were noted with the other four evaluation metrics. The results are illustrated in Figure 4.

Table 6. Fused f-LTR rankers.

Algorithm	Ranker	Understandability		Topicality		
		uRBP	**uRBPgr**	**P@10**	**NDCG@10**	**MAP**
Fused f-LTR	R_{TH_med}	0.7055	0.3000	0.6740	0.6101	0.2492
	R_{THE_med}	0.7220	0.3100	0.7040	0.6246	0.2500
	R_{THEF_med}	**0.7658**	**0.3280**	**0.7440**	**0.6630**	**0.2660**

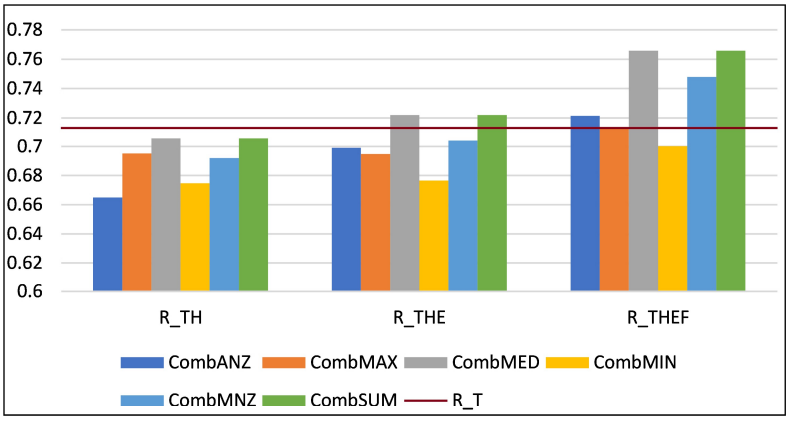

Figure 4. Fusing results from other f-LTR rankers into the title f-LTR ranker using different fusion methods evaluated with the uRBP metric.

As observed, fused f-LTR rankers R_{THEF} were able to surpass the title-based f-LTR ranker R_T with all fusion methods except CombMIN.

CombMIN uses the minimum of the individual similarity values as the combined similarity value. The rationale behind it is to minimize the probability that a non-relevant document would be highly ranked. CombMIN showed worse performance compared to other combination methods, and this is in accordance with previous findings [37].

R_{THEF}_med and R_{THEF}_sum achieved the same and the highest score (0.7658), with an improvement of 7.4% over R_T (0.7131). The results suggest that joining the results obtained from the other f-LTR rankers to the title f-LTR ranker is effective in improving ranking performance.

On the other hand, it can also be observed that when fusing R_H to R_T, the result of R_{TH} was worse in all fusion methods; on the other hand, fusing R_E to R_{TH}, showed improvements over R_T with CombMED and CombSUM fusion methods. Finally, when fusing R_F to R_{THE}, much better scores were achieved when compared to R_T. This demonstrates that not all field information contributes equally, and the way features are explored does affect the ranking performance.

Analyzing the performance of the different fusion methods, all presented similar performance values in all three fusion methods (R_{TH}, R_{THE}, and R_{THEF}), with CombMED and CombSUM achieving the best scores.

5.3. Comparing to the State-of-the-Art Techniques

The results obtained above showed that the proposed f-LTR approach was efficient when compared to the standard LTR and presented much better performance when fused. In this section, the effectiveness of the proposed approach was further examined by comparing it to state-of-the-art techniques.

A group of baselines was built using different IR models and techniques. The best one was chosen for comparison. Eight understandability baselines were built employing state-

of-the-art IR techniques on the PyTerrier platform. Among the eight baselines, six of them were built using TFIDF, BM25, and DirichletLM retrieval models, with or without using pseudo-relevance feedback techniques [38]. The other two baselines were built by using the scores from the reading measures GFI (Gunning Fox Index) and CLI (Coleman–Liau Index).

Six topical relevance baselines were built following the same reasoning used for the understandability baselines: FIDF with and without PRF, BM25 with and without PRF, and DirichletLM with and without PRF. Among them, the TFIDF without PRF baseline was found to be the best one among all baselines and used for comparison.

We also included the eleven runs submitted to the CLEF' 2018 eHealth IR task because they also represented state-of-the-art techniques [39]. Figure 5 presents the results (the baseline, CLEF' 2018's eleven runs, and the f-LTR ranker R_{THEF}_med).

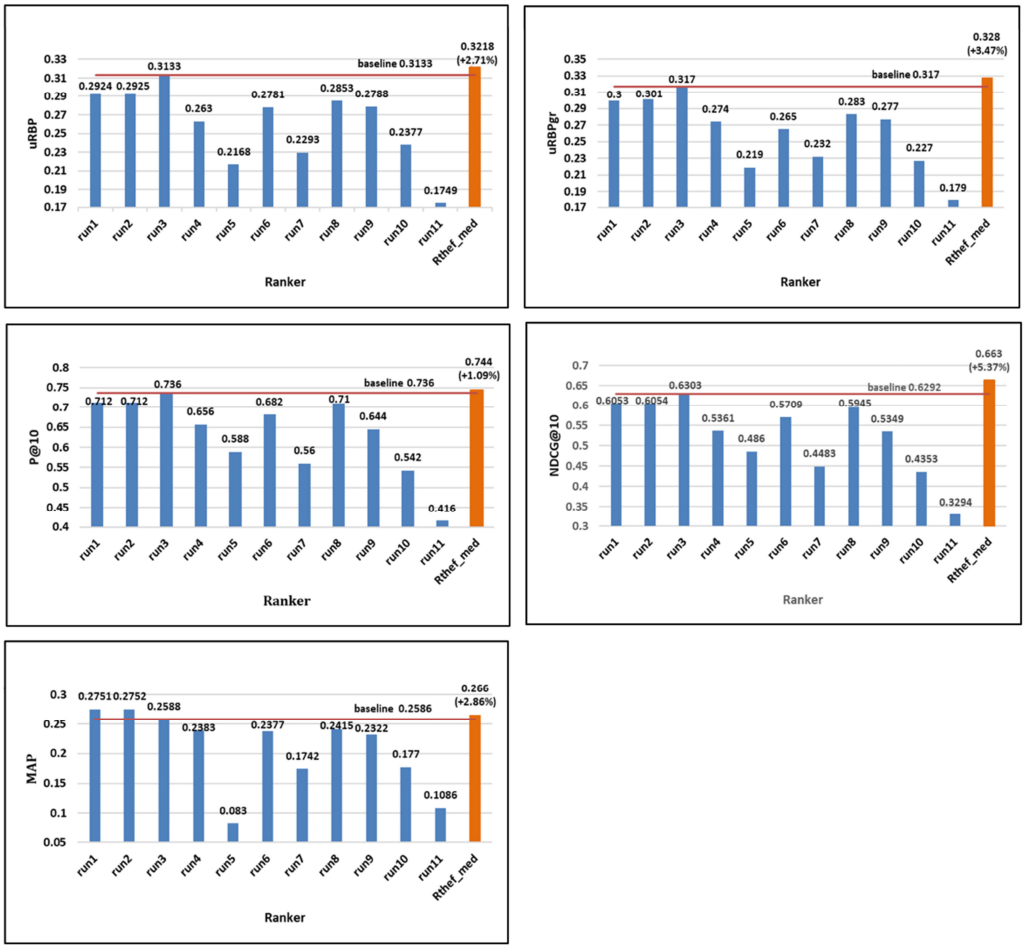

Figure 5. Comparison between the f-LTR approach and state-of-the-art techniques.

Although simple, the baseline was found hard to surpass by the CLEF' 2018 runs, both in understandability (uRBP, uRBPgr) and topical relevance (P@10, NDCG@10, MAP) assessment. Only *run3* was able to achieve similar scores in uRBP, uRBPgr, and P@10 and presented minor improvements in NDCG@10 and MAP; all the other runs failed to exceed the baseline.

By contrast, our ranker R_{THEF}_med exceeded the baseline in all measures with improvements of 2.71% in uRBP, 3.47% in uRBPgr, 1.09% in P@10, 5.37% in NDCG@10, and 2.86% in MAP. Although the improvements were around or under 5%, the built ranker was able to improve the baseline, which, as seen by the CLEF' 2018 runs, was very difficult to surpass. These positive results further prove that the proposed approach is efficient when compared to advanced techniques. Moreover, the obtained results demonstrate that, even

if the existing baseline is strong and hard to exceed, it is possible to build a CHIR system that surpasses it both in terms of understandability and topical relevance.

Further, we compared our method against contemporary competitors in more recent eHealth-related tasks as outlined in CLEF 2024 [40]. The analysis focused exclusively on Mean Average Precision (MAP) results, as this is the only metric reported for the 2024 task. Across all evaluation batches, the median MAP among all participating teams was 0.1311, while the top MAP score was 0.2710; comparatively, our method attained a MAP score of 0.266, placing it close to the highest-performing method. This performance indicates that our approach is competitive with state-of-the-art systems in the eHealth domain, as it is only 0.005 points below the top score. Furthermore, our method significantly exceeds the median performance by 0.135 points, highlighting its efficacy relative to the participating teams [40,41]. These results suggest that our method not only demonstrates robust performance but also contributes in a meaningful way to advancing the state of the art in eHealth-related tasks.

6. Conclusions and Future Work

This paper explored an f-LTR approach to Learning to Rank and its application in the area of consumer health retrieval. The proposed f-LTR approach demonstrated improved results compared to the standard method while utilizing significantly fewer features. Based on the observed results, it can be concluded that the f-LTR approach is an effective solution for enhancing topical relevance and exhibits superior performance concerning understandability assessment in the domain of consumer health services.

The research explored in this paper can be improved and extended in several ways in the future. One valuable finding of this research is that the f-LTR approach is more effective than using all joined features. It would be worthwhile to experiment with additional document-dependent features, such as linguistic information, readability scores, and statistics on medical terminology. Additionally, exploring query-dependent features, such as consumers' readability levels or assessments of their understanding of health information, could yield valuable insights.

Another limitation that should be noted is that this study does not propose a framework for simultaneously optimizing both the relevance and understandability aspects. There is no fusion between the understandability and relevance results, and they are evaluated separately. There may be a conflict between relevance and understandability, and developing a framework to balance these two objectives will be a key focus of future work. We plan to explore multi-objective optimization techniques that can handle such trade-offs, to address the balance between relevance and understandability.

Author Contributions: H.Y.: conceptualization, methodology, data curation, validation, funding acquisition, supervision, writing—original draft. T.G.: methodology, formal analysis, validation, software, resources, funding acquisition, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was supported by Henan Province Key Scientific Research Project Plan for Higher Education Institutions from Henan Provincial Department of Education, namely, the research on the learning-to-rank algorithm by integrating field information and attention mechanism (grant number 24A520060).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed for this study are publicly accessible.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CHS	Consumer health search
CHIR	Consumer health information retrieval
LTR	Learning to Rank
f-LTR	Field-level Learning to Rank
AP	Average Precision
IDCG	Ideal Discounted Cumulative Gain
IR	Information retrieval
MAP	Mean Average Precision
NDCG	Normalized Discounted Cumulative Gain
NIH	National Institute of Health
AMA	American Medical Association
SMOG	Simple Measure of Gobbledygook

References

1. Pugachev, A.; Artemova, E.; Bondarenko, A.; Braslavski, P. Consumer health question answering using off-the-shelf components. In Proceedings of the European Conference on Information Retrieval, Dublin, Ireland, 2–6 April 2023; Springer: New York, NY, USA, 2023; pp. 571–579.
2. Upadhyay, R.; Pasi, G.; Viviani, M. A passage retrieval transformer-based re-ranking model for truthful consumer health search. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Turin, Italy, 18–22 September 2023; Springer: New York, NY, USA, 2023; pp. 355–371.
3. Upadhyay, R.; Knoth, P.; Pasi, G.; Viviani, M. Explainable online health information truthfulness in Consumer Health Search. *Front. Artif. Intell.* **2023**, *6*, 1184851. [CrossRef] [PubMed]
4. Goeuriot, L.; Suominen, H.; Kelly, L.; Alemany, L.A.; Brew-Sam, N.; Cotik, V.; Filippo, D.; Gonzalez Saez, G.; Luque, F.; Mulhem, P.; et al. CLEF eHealth evaluation lab 2021. In Proceedings of the Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, 28 March–1 April 2021; Proceedings, Part II 43; Springer: New York, NY, USA, 2021; pp. 593–600.
5. Zehlike, M.; Castillo, C. Reducing disparate exposure in ranking: A learning to rank approach. In Proceedings of the Web Conference 2020, Taipei, Taiwan, 20–24 April 2020; pp. 2849–2855.
6. Bhatt, C.; Lin, E.; Ferreira-Legere, L.E.; Jackevicius, C.A.; Ko, D.T.; Lee, D.S.; Schade, K.; Johnston, S.; Anderson, T.J.; Udell, J.A. Evaluating readability, understandability, and actionability of online printable patient education materials for cholesterol management: A systematic review. *J. Am. Heart Assoc.* **2024**, *13*, e030140. [CrossRef] [PubMed]
7. Rooney, M.K.; Santiago, G.; Perni, S.; Horowitz, D.P.; McCall, A.R.; Einstein, A.J.; Jagsi, R.; Golden, D.W. Readability of patient education materials from high-impact medical journals: A 20-year analysis. *J. Patient Exp.* **2021**, *8*, 2374373521998847. [CrossRef] [PubMed]
8. Deidra Bunn, S.; Erickson, K. Voices from Academia Minimizing the Complexity of Public Health Documents: Making COVID-19 Documents Accessible to Individuals Who Read Below the Third-Grade Level. In *Assistive Technology Outcomes and Benefits Accessible Public Health Materials During a Pandemic: Lessons Learned from COVID-19*; Assistive Technology Outcomes & Benefits (ATOB): Schaumburg, IL, USA, 2022.
9. Kher, A.; Johnson, S.; Griffith, R. Readability assessment of online patient education material on congestive heart failure. *Adv. Prev. Med.* **2017**, *2017*, 9780317. [CrossRef]
10. Hollada, J.L.; Zide, M.; Speier, W.; Roter, D.L. Readability Assessment of Patient-Centered Outcomes Research Institute Public Abstracts in Relation to Accessibility. *Epidemiology* **2017**, *28*, e37–e38. [CrossRef]
11. Antunes, H.; Lopes, C.T. Proposal and comparison of health specific features for the automatic assessment of readability. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Xi'an, China, 25–30 July 2020; pp. 1973–1976.
12. Gordejeva, J.; Zowalla, R.; Pobiruchin, M.; Wiesner, M. Readability of English, German, and Russian Disease-Related Wikipedia Pages: Automated Computational Analysis. *J. Med. Internet Res.* **2022**, *24*, e36835. [CrossRef]
13. Liu, T.Y. Learning to rank for information retrieval. *Found. Trends® Inf. Retr.* **2009**, *3*, 225–331. [CrossRef]
14. Burges, C.J. From ranknet to lambdarank to lambdamart: An overview. *Learning* **2010**, *11*, 81.
15. Joachims, T. Optimizing search engines using clickthrough data. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–26 July 2002; pp. 133–142.
16. Miyachi, Y.; Ishii, O.; Torigoe, K. Design, implementation, and evaluation of the computer-aided clinical decision support system based on learning-to-rank: Collaboration between physicians and machine learning in the differential diagnosis process. *BMC Med. Inform. Decis. Mak.* **2023**, *23*, 26. [CrossRef]

17. Javaid, M.; Haleem, A.; Singh, R.P.; Suman, R.; Rab, S. Significance of machine learning in healthcare: Features, pillars and applications. *Int. J. Intell. Netw.* **2022**, *3*, 58–73. [CrossRef]
18. Habehh, H.; Gohel, S. Machine learning in healthcare. *Curr. Genom.* **2021**, *22*, 291. [CrossRef] [PubMed]
19. Geng, X.; Liu, T.Y.; Qin, T.; Li, H. Feature selection for ranking. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, 23–27 July 2007; pp. 407–414.
20. Xu, J.; Li, H. Adarank: A boosting algorithm for information retrieval. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, 23–27 July 2007; ACM: New York, NY, USA, 2007; pp. 391–398.
21. Douze, L.; Pelayo, S.; Messaadi, N.; Grosjean, J.; Kerdelhué, G.; Marcilly, R. Designing Formulae for Ranking Search Results: Mixed Methods Evaluation Study. *JMIR Hum. Factors* **2022**, *9*, e30258. [CrossRef] [PubMed]
22. Azarbonyad, H.; Dehghani, M.; Marx, M.; Kamps, J. Learning to rank for multi-label text classification: Combining different sources of information. *Nat. Lang. Eng.* **2021**, *27*, 89–111. [CrossRef]
23. Ueda, A.; Santos, R.L.; Macdonald, C.; Ounis, I. Structured Fine-Tuning of Contextual Embeddings for Effective Biomedical Retrieval. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual, 11–15 July 2021; pp. 2031–2035.
24. Fox, E.A.; Shaw, J.A. Combination of multiple searches. In Proceedings of the 2nd Text REtrieval Conference (TREC-2), Gaithersburg, MD, USA, 31 August–2 September 1993; NIST Special Publication 500-215; pp. 243–252.
25. Vogt, C.C.; Cottrell, G.W. Fusion via a linear combination of scores. *Inf. Retr.* **1999**, *1*, 151–173. [CrossRef]
26. Manmatha, R.; Rath, T.; Feng, F. Modeling score distributions for combining the outputs of search engines. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, LA, USA, 9–13 September 2001; ACM: New York, NY, USA, 2001; pp. 267–275.
27. Kuzi, S.; Shtok, A.; Kurland, O. Query expansion using word embeddings. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, Indianapolis, IN, USA, 24–28 October 2016; ACM: New York, NY, USA, 2016; pp. 1929–1932.
28. Xia, X.; Lo, D.; Wang, X.; Zhang, C.; Wang, X. Cross-language bug localization. In Proceedings of the 22nd International Conference on Program Comprehension, Hyderabad, India, 2–3 June 2014; ACM: New York, NY, USA, 2014; pp. 275–278.
29. Ru, X.; Ye, X.; Sakurai, T.; Zou, Q. Application of learning to rank in bioinformatics tasks. *Briefings Bioinform.* **2021**, *22*, bbaa394. [CrossRef]
30. Bhagawati, R.; Subramanian, T. An approach of a quantum-inspired document ranking algorithm by using feature selection methodology. *Int. J. Inf. Technol.* **2023**, *15*, 4041–4053. [CrossRef]
31. Zhao, Y.; Da, J.; Yan, J. Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. *Inf. Process. Manag.* **2021**, *58*, 102390. [CrossRef]
32. Oyebode, O.; Fowles, J.; Steeves, D.; Orji, R. Machine learning techniques in adaptive and personalized systems for health and wellness. *Int. J. Hum. Comput. Interact.* **2023**, *39*, 1938–1962. [CrossRef]
33. Henrich, A.; Wegmann, M. Search and evaluation methods for class level information retrieval: extended use and evaluation of methods applied in expertise retrieval. In Proceedings of the 36th Annual ACM Symposium on Applied Computing, Gwangju, Republic of Korea, 22–26 March 2021; pp. 681–684.
34. Macdonald, C.; Tonellotto, N.; MacAvaney, S.; Ounis, I. PyTerrier: Declarative experimentation in Python from BM25 to dense retrieval. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Gold Coast, Australia, 1–5 November 2021; pp. 4526–4533.
35. Aloteibi, S. A User-Centred Approach to Information Retrieval. Ph.D. Thesis, University of Cambridge, Cambridge, UK, 2021.
36. Santos, P.M.; Teixeira Lopes, C. Generating query suggestions for cross-language and cross-terminology health information retrieval. In Proceedings of the Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, 14–17 April 2020; Proceedings, Part II 42; Springer: New York, NY, USA, 2020; pp. 344–351.
37. Bałchanowski, M.; Boryczka, U. How Normalization Strategies Affect the Quality of Rank Aggregation Methods in Recommendation Systems. *Procedia Comput. Sci.* **2023**, *225*, 1843–1852. [CrossRef]
38. Azad, H.K.; Deepak, A. Query expansion techniques for information retrieval: A survey. *Inf. Process. Manag.* **2019**, *56*, 1698–1735. [CrossRef]
39. Jimmy; Zuccon, G.; Palotti, J. Overview of the CLEF 2018 Consumer Health Search Task. In Proceedings of the Working Notes of Conference and Labs of the Evaluation (CLEF) Forum, CEUR Workshop Proceedings, Avignon, France, 10–14 September 2018.
40. Nentidis, A.; Katsimpras, G.; Krithara, A.; Paliouras, G. Overview of BioASQ tasks 12b and Synergy12 in CLEF2024. *Work. Notes CLEF* **2024**, *2024*. Available online: https://ceur-ws.org/Vol-3740/paper-01.pdf (accessed on 28 October 2024).
41. Şerbetçi, O.; Wang, X.D.; Leser, U. HU-WBI at BioASQ12B Phase A: Exploring Rank Fusion of Dense Retrievers and Re-rankers. In Proceedings of the Conference and Labs of the Evaluation Forum, Grenoble, France, 9–12 September 2024.