Information Systems 121 (2024) 102315

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is

Document structure-driven investigative information retrieval

Tuomas Ketola [*], Thomas Roelleke

Queen Mary University of London, United Kingdom

ARTICLE INFO

Recommended by Dennis Shasha

Keywords:
Structured document retrieval
Transparent methods
Analytical models
Investigative search

ABSTRACT

Data-driven investigations are increasingly dealing with non-moderated, non-standard and even manipulated information Whether the field in question is journalism, law enforcement, or insurance fraud it is becoming more and more difficult for investigators to verify the outcomes of various black-box systems To contribute to this need of discovery methods that can be used for verification, we introduce a methodology for document structure-driven investigative information retrieval (InvIR) InvIR is defined as a subtask of exploratory IR, where transparency and reasoning take centre stage The aim of InvIR is to facilitate the verification and discovery of facts from data and the communication of those facts to others From a technical perspective, the methodology applies recent work from structured document retrieval (SDR) concerned with formal retrieval constraints and information content-based field weighting (ICFW) Using ICFW, the paper establishes the concept of relevance structures to describe the document structure-based relevance of documents These contexts are then used to help the user navigate during their discovery process and to rank entities of interest The proposed methodology is evaluated using a prototype search system called Relevance Structure-based Entity Ranker (RSER) in order to demonstrate its the feasibility This methodology represents an interesting and important research direction in a world where transparency is becoming more vital than ever.

1. Introduction

Investigations focused on large data collections are increasingly dealing with non-moderated, non-standard, messy and even manipulated data In fields such as journalism, law-enforcement, insurance fraud and open-source investigations, it is becoming more difficult to verify the outcomes of various black-box systems To contribute to the issue of verification in such data-driven investigations (DDIs), this paper proposes a new methodology for structure-driven Investigative Information Retrieval (InvIR) and evaluates the viability of this new methodology using a prototype discovery system called Relevance Structure-based Entity Ranker (RSER).

When describing the field of InvIR, we identify a set of features, which the underlying retrieval models must possess, in order for them to be useable in the field.

1. The models must leverage *structure*.
2. The models must have a high level of *transparency*.
3. The models must be *analytical* in nature.

When discussing potential existing IR methods applicable to InvIR, the paper recognises recent research on structured document retrieval (SDR) constraints and the Information Content-based Field Weighting (ICFW) technique as important methods for enhancing retrieval and navigation in InvIR Together these two lines of research provide the necessary transparency and analytical capabilities required for InvIR.

The proposed methodology and prototype is evaluated on two data-collections which have been extended from two well known IR benchmarks: DBpedia and IMDB The evaluation demonstrates that by using the ICFW method, RSER is able to leverage the document structures for easier navigation, better performance and importantly for communicating the context in which entities occur in the data to the user The aim of the evaluation is not to show that RSER can automate InvIR, but to demonstrate and analyse the validity and feasibility of the proposed methodology.

The rest of the paper is structured as follows: Section 2 introduces InvIR and describes what is required of the underlying IR models used in it, Section 3 is an in-depth analysis of existing analytical structured document retrieval (SDR) models and how well they are suited for InvIR, Section 4 discusses the concept of relevance structures and how existing models relate to it, Section 5 details the RSER system, Section 6 presents an evaluation of the system, Section 7 discusses how the proposed methodology in a wider context and Section 8 summarises and concludes.

[*] Corresponding author.
E-mail address: t.j.h.ketola@qmul.ac.uk (T. Ketola).

https://doi.org/10.1016/j.is.2023.102315
Received 27 September 2023; Received in revised form 6 November 2023; Accepted 8 November 2023
Available online 19 November 2023

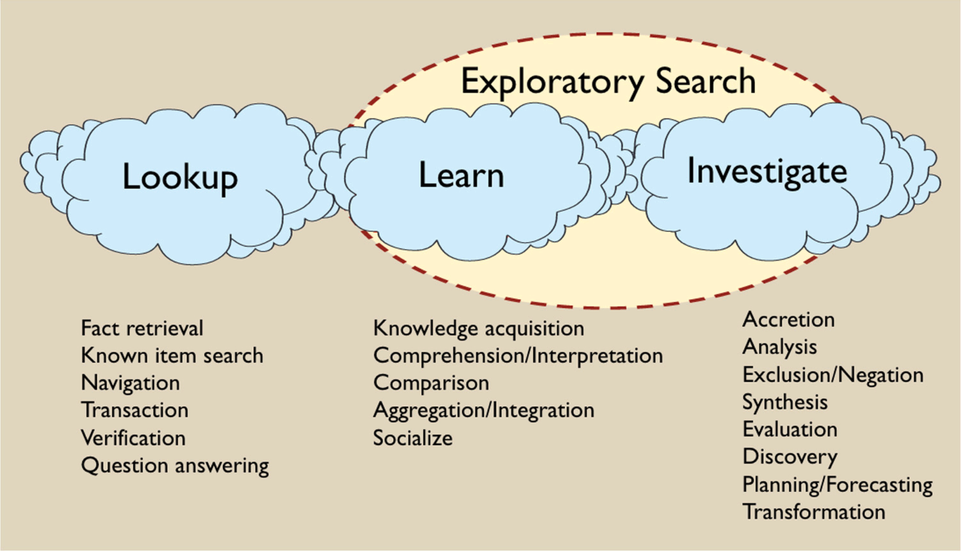


Fig. 1. Categorisation of search activities by [1]. Activities relating to InvIR furthest to the right and ExIR activities not as related to investigations in the middle.

2. Investigative Information Retrieval (InvIR) and the need for analytical and transparent models

This section introduces the concept of InvIR as a type of exploratory search; or exploratory information retrieval (ExIR) First, it is worth clarifying some of the terminology Exploratory search refers to any search activity where a user is looking to learn something new from the data, rather than simply looking for factual answers to questions This means that their information needs tend to be much more complicated, meaning often their search session would be much more complex and long [1] Fig. 1 shows the different kinds of search tasks that relate to exploratory search ExIR as a term is not used as widely as exploratory search Here, the two are used interchangeably.

InvIR is similar to ExIR, in fact, it can be seen as a subset of it InvIR refers to any search task where the goal is to learn new facts from the data that are interesting, not only to the user, but to other people as well The difference between exploratory search and InvIR is the emphasis of the latter on facts and other people A simple example of an InvIR scenario would be an investigative journalist searching a for new stories and facts in data Their ultimate aim is to communicate what they find to readers and for those readers to believe them.

The emphasis on other people has implications in terms of what is required of the retrieval methods used in InvIR In order for the information found to be reportable by a journalist for example, it has to be believable, meaning the journalist has to understand what they have found, how they have found it and how it can be trusted Furthermore, they must be able to communicate all of this to their reader This means that an InvIR system has to be transparent and a user has to be able to reason with the system in order to understand all of the relevant information Effectively this means that all the facts, or outcomes that the investigative system has produced have to be verifiable, and the system should facilitate this verification.

Table 1 illustrates how ad-hoc IR, exploratory IR and InvIR are related It relates to Fig. 1 in that search activities shown there can be seen in the light of the three types of IR Ad-hoc IR deals with more straightforward information needs and can therefore be seen as a "lookup" search activity Exploratory IR covers both learning and investigating search InvIR can be seen to specifically relate to the rightmost search activities Table 1 looks at the various aspects that make InvIR a sub-category of exploratory search From the table it is evident that what separates InvIR from ExpIR is the emphasis on transparency and reasoning.

Table 1
Differences and similarities between Ad-hoc IR, Exploratory IR and InvIR. The emphasis on transparency and reasoning is what differentiates InvIR from Exploratory IR.

Aspect	Ad-hoc IR	Exploratory IR	Investigative IR
Complex Info. Needs	Optional	Essential	Essential
Query Reformulation	Optional	Essential	Essential
Session-based	Optional	Essential	Essential
Complex Results	Optional	Essential	Essential
Complex Data	Optional	Essential	Essential
Transparency	Optional	Optional	Essential
Reasoning	Optional	Optional	Essential

Data-driven investigations (DDIs) deal with relatively complex collections, with structured data such as emails, legal contracts, spreadsheets and company registers playing a central part For this reason, retrieval models used in InvIR, should be able to leverage document structures Hence the methodology proposed here is structure-driven.

In conclusion, from the vast array of existing IR models and approaches we identify analytical SDR models as the subset on which we focus This is because they are transparent and they are able to leverage document structures Their analytical nature also facilitates reasoning, as the user can transparently understand what the model is doing and therefore they can reason with it.

3. Analytical structured document retrieval (SDR)

This section starts by clarifying the distinction between analytical and non-analytical retrieval models It then summarises recent research on formal constraints for SDR and how that research contributes to retrieval transparency Finally, it summarises the ICFW method and the features that make it a strong potential underlying model for InvIR systems.

3.1. Analytical vs non-analytical retrieval

This paper considers a model analytical if its ranking behaviour can be inferred from its specification without further knowledge. For example given a query, two documents and a retrieval model (with known hyperparameters), if the model is analytical the ranking of the documents can be inferred without having to process the documents, or query in any way. A non-analytical model would be one where such inference is not possible, such as those involving large language models (LLMs).

Table 2
Intuition underlying formal constraints for SDR. Field refers to a field of a document; e.g. *abstract* or *author*. Table from [3] with an additional cell for Term Importance.

	Term	Field
Importance	A model should consider the importance of a term on a field level, rather than document-level	A model should be able to boost, or decrease the weight given to a field-based on some notion of field importance
Distinctiveness	Adding unseen query terms to a document should increase the retrieval score more than adding query terms already considered	Adding a query term to a new field should increase the retrieval score more than adding it to a field where it already occurs

Analytical models (e.g. BM25, DFR, LM etc.) are often considered to be less powerful than their supervised learning-based counterparts, the reason being that they do not leverage training data to the same extent, or in the same manner In essence the learning procedure for analytical models does not seek to bridge the semantic gap, which is the main reason why non-analytical models perform better when training data is available However, analytical models have three important advances over non-analytical models:

- They are more robust across different data collections, especially when separate training data is not available for each collection.
- They are more transparent, meaning their inner workings are easy to dissect. This is due to them having a more solid theoretical grounding amongst other things [2].
- They tend to be faster as the retrieval scores can be calculated directly from the index.

The first two points are crucial in the context of this paper, as transparency and lack of training data are distinguishing characteristics of an InvIR scenario, which is why the focus in this paper is on analytical models.

3.2. Formal constraints for SDR

Ketola et al. introduced the formal constraints for SDR [3] Like their counterparts in atomic retrieval, they can be used to provide an additional layer of transparency to any analytical SDR model by dissecting how the model would behave in certain situations This is important for retrieval models used by the RSER system introduced later in this paper, as InvIR requires a high degree of transparency Table 2 from summarises the intuition behind the four SDR constraints. A fourth constraint for Term Importance was added by [4] This constraint was added because, a term might carry a different meaning depending on the field it occurs in, and therefore its occurrences in different fields should be treated separately.

3.3. Information Content Field Weighting (ICFW)

[4] introduced the Information Content Field Weighting method Its underlying idea is to emphasise document fields which carry more information, just as the TF-IDF emphasises terms which are rarer and therefore carry more information Definition 3.1 shows how ICFW calculates the retrieval score See [4] for a more in-depth description of the model.

Definition 3.1 (*ICFW Retrieval Score*)**.** Let S_M be a retrieval score of retrieval model M (e.g. BM25), ICD be the document field based information content and ICF the collection field based information content The scaling parameter λ controls the weight given to the document field based information content and thus the degree of term frequency saturation across fields. Given document d, query q, collection c and retrieval model M, the score (retrieval status value) of d is denoted $\mathrm{RSV}_{\mathrm{ICFW},\lambda,M}(d,q,c)$.

$$\mathrm{RSV}_{\mathrm{ICFW},\lambda,M}(d,q,c) := \sum_{i=1}^{m} w_{\mathrm{icfw},\lambda_i}(f_i,F_i,d,q)\,\mathrm{RSV}_M(q,f_i,c) \tag{1}$$

□

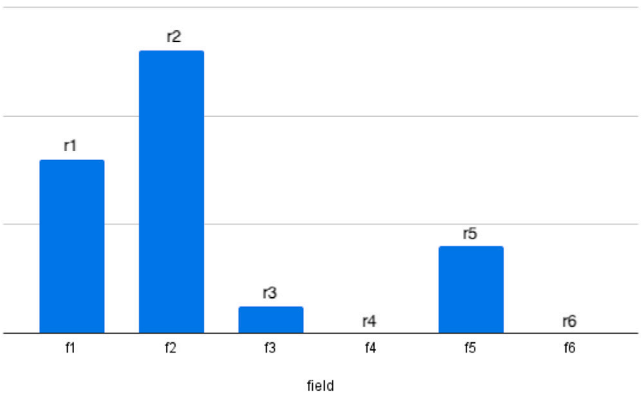


Fig. 2. The relevance structure of document d ($d = [f_1 \dots f_6]$) presented as a bar graph.

The reason for the focus on ICFW, is that it is the only existing analytical SDR model that satisfies all the retrieval constraints from Section 3.2, as can be seen from Table 3, meaning it is likely to perform better than other analytical SDR models The next section discusses other reasons for why ICFW is expected to work better in terms of the task at hand.

4. From document structures to relevance structures

The ICFW method leverages document structures to increase retrieval performance One of the aims of this paper is to take the ICFW assigned field weights ($w_{\mathrm{icfw},\lambda_i}(f_i,F_i,d,q)$) and use them to help the user navigate the data and to better understand how the documents structures affect relevance This is where the concept of relevance structures comes in.

A relevance structure describes how the structure of the document contributes to the system's perceived relevance of a document with respect to a query, i.e. the retrieval score Put in another way, relevance structure describes the composition of a document's relevance with respect to its fields Relevance structure can be visualised in many ways Fig. 2 demonstrates the use of a histogram, which is also used in the prototype system More formally, relevance structure is defined as follows:

Definition 4.1 (*Relevance Structure*)**.** Let r_i be the relevance of a document field f_i The relevance structure vector of a document with respect to query q and collection c is denoted $\vec{\mathrm{rs}}$

$$\vec{\mathrm{rs}}(q,d,c) := \left[r_1(q,f_1,c)\dots r_m(q,f_i,c)\right] \tag{2}$$

□

r_i can be defined in many different ways The most naive method would be to define it as a field-based RSV: $r_i(q,f_i,c) := \mathrm{RSV}_M(q,f_i,c)$ However, as was shown by [3,4] raw field-based scores do not model the relevance of document fields well if there is significant dependence between term occurrences across the fields This means that the relevance structure vectors could become very noisy ICFW was developed for exactly this purpose; to saturate term frequency across fields in order to model the dependence of term occurrence across fields Therefore, it should be much better in estimating the r_i values in Definition 4.1.

Table 3
Constraint satisfaction of SDR models, including ICFW Table extended from [3] Conditional satisfaction of a constraint refers to cases where collection statistics need to be accounted for, i.e. the specificity/IDF of query terms for example FSA: Field Score Aggregation, PRMS: Probabilistic Model for Semi-structured data, MLM: Mixture of Language models, FSDM: Fielded Sequential Dependence Model.

	Term Distinct. TD-Co	Field Distinct. FD-Co	Term Import. TI-Co	Field Import. FI-Co
FSA [3]	NO	Conditional	YES	YES
BM25-FIC [5]	NO	Conditional	YES	YES
PRMS [6]	NO	Conditional	NO	YES
BM25F [7]	Conditional	NO	NO	YES
MLM [8]	Conditional	NO	NO	YES
FSDM [9]	Conditional	NO	NO	YES
ICFW [4]	**Conditional**	**Conditional**	**YES**	**YES**

5. Relevance Structure-based Entity Ranker (RSER)

This section presents a prototype InvIR system called Relevance Structure-based Entity Ranker (RSER) The system is built to demonstrate and investigate the methodology of structure-driven transparent InvIR proposed in this paper RSER uses relevance structures to define a context in which a seed entity occurs in a dataset and using this context ranks other entities of interest (EoIs) according to whether they are found in a similar context The prototype in its entirety, as well as the evaluation, is available at https://github.com/TuomasKetola/relevance-structure-ranker.

5.1. Entity ranking based on context and relevance structures

Search is a central aspect of DDIs where investigators dig through data collections for previously unknown facts. These users can be journalists exploring public data and leaks,[1][2] law enforcement offices investigating data obtained through foreclosures, or open source investigators scouring social media data for evidence of dubious activity etc [10,11]. In all the scenarios above, it is likely that the investigators have a list of “Entities of Interest” (EoIs) that they think could be found in the data. Furthermore, they might already know of an interesting entity in the data, this “seed entity” (SE) can be used as a reference point.

More formally, retrieval task above can be described as follows: Given a user’s information need, a list of EoIs, the user’s knowledge of a SE and their specific interest in the SE, rank the list of EoIs based on whether they can be found in a similar context as the SE in the data.

As a more concrete example consider the following: The information need is “List of Russian people that keep money in tax heavens, own a yacht and have ties to the government?”, the SE could be Arkady Rotenberg, who is known to satisfy the information need well. The list of EoIs could be every influential person in Russia for example (n=10k+) Given this information we would like to rank the EoIs based on whether they can be found in a similar context in the data as Arkandy Rotenberg, i.e. whether they have money in tax heavens, own a yacht and have ties to the government This would significantly ease the work of the investigator, as they would have a better idea of which entities they should start with It is imperative that a user can easily understand the inner workings of the system in terms of why it produces the ranking it does Otherwise, the investigator cannot trust the system This is why so much emphasis is given to the transparency of the system.

The above example describes the motivation for the proposed system well in the context of investigations and the reduction of labour for the investigator However, its complexity makes it difficult to clearly explain the inner workings of the system For this purpose, it is easier to consider an example with movie-related data Table 4 demonstrates how the engine defines the context and how the EoIs are ranked for data about movies, actors and characters.

Table 4
Example entity ranking scenario where a user is looking to rank entities in movies about magic based on context q = *wizards magic fantasy* First context is actors, second context is characters SE = seed entity SE$_1$ = Emma Watson, SE$_2$ = Hermione.

rank	ranking for SE$_1$	ranking for SE$_2$
1	Tom Felton	Malfoy
2	Nicole Kidman	Bilbo Baggins
3	Robin Williams	Alladin
4	Martin Freeman	Coulter
5	Coulter	Tom Felton
6	Malfoy	Nicole Kidman
7	Alladin	Martin Freeman
8	Bilbo Baggins	Robin Williams

The list of entities contains actors and characters from movies about magic: Malfoy, Bilbo Baggins, Alladin, Coulter, Tom Felton, Nicole Kidman, Martin Freeman, and Robin Williams Consider two information needs, the first one is actors in movies about magic and the second one is characters in movies about magic The information need is described by the base query “wizards magic fantasy” and a seed entity that is chosen by the user For the first information need the user chooses Emma Watson who they know is an actor in a magic movie (Harry Potter) and for the second one they choose Hermione as they know she is a character Given this information, the system, with the help of the user, should produce the two rankings of EoIs in Table 4 corresponding to the two contexts.

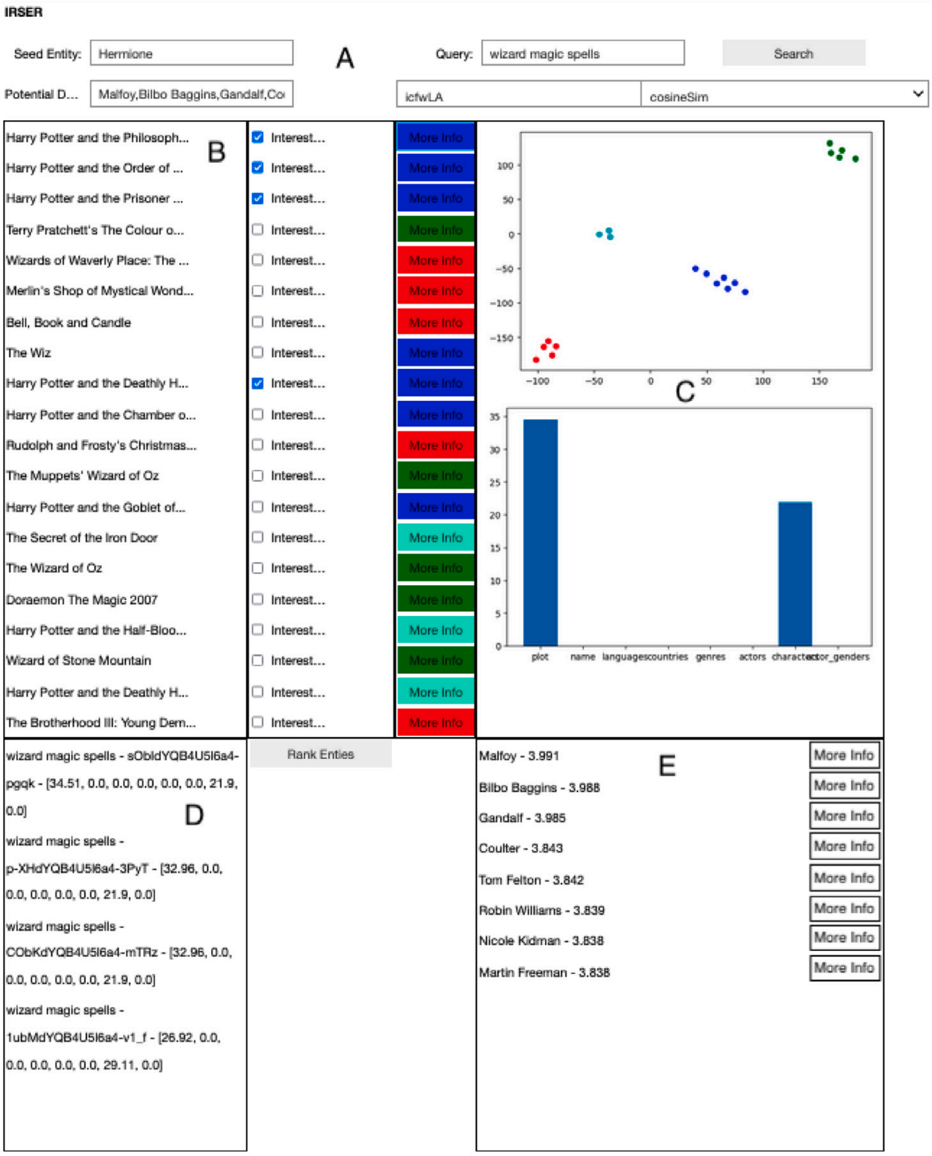
5.2. System description

Before diving into the technical description of the system described in this section, it is worth clarifying the notation used:

- SE: a seed entity, i.e. someone we know matches the information need well and can be found in the data
- EoIs = (pe$_1$ … pe$_m$): a list of potential entities of interest who might match the information need
- $Q(e) = (q_1(e) \dots q_m(e))$: a set of queries. Each query q contains an entity e. This can be either the seed entity or one of the potential entities. For SE= Hermione in Table 4 the q(Hermione) = “wizards magic fantasy Hermione”.
- SIM$[\mathrm{rs}(q_i, d_i), \mathrm{rs}(q_j, d_j), \gamma]$: a function that returns the similarity of two relevance structures. γ denotes a chosen similarity model

Fig. 3 shows the RSER user interface (UI), which will be used to describe how the system works in detail The example is based on a simple scenario where the data collection consists of information about movies, actors and characters The reason for the simple scenario is to make it easier to follow, but there is no reason why the scenario could not involve more interesting and complex data Altogether there are six steps in the process The following details each of these steps.

[1] https://aleph.occrp.org/
[2] https://datashare.icij.org/



Seed Entity: Hermione

Potential D... Malfoy,Bilbo Baggins,Gandalf,Co

Query: wizard magic spells

icfwLA

cosineSim

☑ Harry Potter and the Philosoph... Interest...
☑ Harry Potter and the Order of ... Interest...
☑ Harry Potter and the Prisoner ... Interest...
□ Terry Pratchett's The Colour o... Interest...
□ Wizards of Waverly Place: The ... Interest...
□ Merlin's Shop of Mystical Wond... Interest...
□ Bell, Book and Candle Interest...
□ The Wiz Interest...
☑ Harry Potter and the Deathly H... Interest...
□ Harry Potter and the Chamber o... Interest...
□ Rudolph and Frosty's Christmas... Interest...
□ The Muppets' Wizard of Oz Interest...
□ Harry Potter and the Goblet of... Interest...
□ The Secret of the Iron Door Interest...
□ The Wizard of Oz Interest...
□ Doraemon The Magic 2007 Interest...
□ Harry Potter and the Half-Bloo... Interest...
□ Wizard of Stone Mountain Interest...
□ Harry Potter and the Deathly H... Interest...
□ The Brotherhood III: Young Dem... Interest...

wizard magic spells - sObldYQB4U5I6a4-pgqk - [34.51, 0.0, 0.0, 0.0, 0.0, 0.0, 21.9, 0.0]

wizard magic spells - p-XHdYQB4U5I6a4-3PyT - [32.96, 0.0, 0.0, 0.0, 0.0, 0.0, 21.9, 0.0]

wizard magic spells - CObKdYQB4U5I6a4-mTRz - [32.96, 0.0, 0.0, 0.0, 0.0, 0.0, 21.9, 0.0]

wizard magic spells - 1ubMdYQB4U5I6a4-v1_f - [26.92, 0.0, 0.0, 0.0, 0.0, 0.0, 29.11, 0.0]

Malfoy - 3.991
Bilbo Baggins - 3.988
Gandalf - 3.985
Coulter - 3.843
Tom Felton - 3.842
Robin Williams - 3.839
Nicole Kidman - 3.838
Martin Freeman - 3.838

Fig. 3. User interface for RSER demonstrating the main components. A: User input, B: Ranking results and choosing documents of interest, C: Graphic explanation of results, D: Chosen documents of interest, E: Final entity ranking.

5.2.1. Step 1: Defining the entities of interest

As with any search task, the investigatory process begins with the user having an information need In the proposed system, the user defines a list of potential entities (EoIs) they are interested in and believe could be found in the data Furthermore, the user defines a seed entity (SE) that they know is found in the data and corresponds to their information need.

In the example from Table 4 and Fig. 3 the list of potential entities is a mix of character and actor names from movies with magic and wizards: EoIs = [Malfoy, Bilbo Baggins, Alladin, Coulter, Tom Felton, Nicole Kidman, Martin Freeman, Robin Williams]. For Fig. 3 and the explanation of the process the seed entity SE = Hermione This suggests that the user is interested in characters found in magical movies, rather than actors.

5.2.2. Step 2: Formulating a set of queries

This step consists of the user formulating a set of queries Q that define the context in which they wish to rank the list of potential entities In our movie example, this context is whether an entity is a character in a movie about magic and wizards To define this context — with the help of the seed entity — the user should formulate queries that return documents proving that Hermione is a character in such a movie.

These queries should not be too specific, as something like "Harry Potter Hermione" would return the Harry Potter movies, but the associated relevance structures would carry little information in the context of most of the potential entities Something like "wizards magic spells Hermione" would give better results The user can also choose the retrieval model and the similarity model in section A of the user interface The different possible models are discussed in Section 6.

5.2.3. Step 3: Choosing documents of interest

On the left of Fig. 3 (Section B) we can see the produced ranking To the right of it are some graphs that help navigate the ranking (Section C) The top one shows the documents on the left clustered based on their relevance structures The clustering is performed based on the relevance structure of the documents, i.e. the field weights TSNE clustering is used with the number of clusters calculated using Silhouette Coefficients.

The user can easily examine the relevance structure of each document by clicking the "More info" button which updates the bottom graph From this graph, the user can get an understanding of how the document is relevant to the query in terms of its structure See Fig. 2 In this case, we can clearly see that the document is relevant because of the plot and actor fields.

Using the cluster colours and the bottom graphs the user can easily navigate the results based on their relevance structures In Fig. 3 this is clear from the Harry Potter movies all being blue The user can easily select interesting documents which then appear in the list at the bottom left of the UI They can re-run the query and add more items to the list Once they believe they have enough interesting documents the user simply clicks "Rank Entities", which starts the back-end analysis of the EoIs.

5.2.4. Step 4: Calculating similarity scores for EoIs

This step is performed by the system, not the user For each entity in the list of EoIs, we define a set of queries Q(entity) and for each of these queries, we run a search So for example, for the entity "Malfoy" we would run the query "wizards magic spells Malfoy", just as we ran 'wizards magic spells Hermione" for the seed entity We would then look at the ranking produced and see if there are relevance structures similar to those defined in the previous step, i.e. the interesting documents were chosen.

The similarity of each of the document relevance structures in the ranking for "wizards magic spells Malfoy" would be compared to those chosen by the user for the seed entity, using the similarity model $\mathrm{SIM}[\mathrm{rs}(q_i, d_i), \mathrm{rs}(q_j, d_j), \gamma]$ where d_i is a document of interest chosen by the user, and d_j is a document in the ranking corresponding to the query "wizards magic spells Malfoy" To calculate the similarity for a given EoI, we consider k most similar documents from the EoIs rankings, compared to the documents of interest Section 6 will discuss different options for measuring this similarity and the performance for different values of k.

5.2.5. Step 5: Rank the potential entities according to their relevance to the information need

Create the final ranking and allow the user to investigate the underlying queries, documents and fields which have produced said ranking This is done by sorting the entities based on the SIM scores that were calculated in the previous step.

5.2.6. Step 6: Reasoning

If the user wishes to learn about the system's reasoning for why an EoI gets a certain similarity score they can click the "More Information" button to the right of each EoI This will show the ranking results for that specific EoIs queries in the main ranking table, as well as the documents of interest that are the reason for its similarity score By deleting documents of interest that are not relevant to their information needs, the investigator can then "reason" with the system.

5.3. Implementation of RSER

The proposed RSER system has been implemented using ElasticSearch, python and Jupyter notebooks Elasticsearch is used to store the data and to perform the initial field-based queries using BM25 and BM25F For each field in the data collection, we retrieve the top 1000 documents with BM25 Furthermore, we retrieve the top 1000 documents with the BM25F using all the fields and calculate their field-based scores Hyperparameters are set as $b = 0.8$ and $k_1 = 1.6$ A python library re-ranks the retrieved documents and calculates their field weights using ICFW We test the system using all three proposed ICFW versions [4].

A Jupyter notebook is used to create the user interface (UI) presented in Fig. 3 The UI is not a part of the evaluation The system is only evaluated on its performance in terms of a benchmark test collection created specifically for this paper.

5.4. RSER and InvIR

Before moving onto the evaluation of the proposed system, it is worth discussing what exactly makes RSER an investigative search engine, rather than just a search engine, or an exploratory search engine.

From Table 1 complex queries, query reformulation and the session-based nature of RSER are evident from the previous section The system considers multiple queries, each of which is comprised of two parts (base query + entity) within a session where the end result is to rank entities of interest The results are presented in a complex manner, where the relevance structures are explorable both one by one and from clusters To the best of the author's knowledge, this is the first system that visualises the relationship between relevance and document structures (relevance structures) in this manner It makes the search much easier as effectively documents with similar relevance structures, i.e. documents that are relevant in the same context have the same colour in the ranking The complexity of data is also evident, as the search engine is specifically designed to deal with structured data.

Moving onto the aspects of RSER that make it an investigative search engine, rather than an exploratory one The emphasis on transparency does not have so much to do with the engine design itself, but rather with the underlying retrieval and similarity models If either of these was replaced by a black-box algorithm a large degree of the transparency of the system would be lost, which would make it an exploratory search engine instead A degree of transparency is also provided by the way in which the results are presented, as the user has a better idea of which parts of the document contribute to its relevance.

Even if the underlying algorithms are transparent, a user with little knowledge about IR algorithms cannot fully trust the system That is why the ability of the system to communicate the reasoning of the final ranking of the EoIs to the user and the users' ability to reason together with the system is an important aspect of what makes RSER an investigative search engine This is why it is important that the user can get a deeper understanding of why each of the EoIs has been ranked high or low The system accomplishes this by allowing the user to see the rankings that each of the queries for each EoI entity has produced and the documents of interest the system has chosen for those query–entity pairs If the documents of interest are not correct, the user should be able to change them, effectively reasoning together with the system to change the final ranking.

Analysing the existing body of research around reasoning as a cooperative and interactive task between a human and a computer in-depth is out of the scope of this paper For example, there exist whole fields of study around concepts such as semantic web, semantic reasoning and reasoning based on knowledge bases that relate to this chapter but are too wide to capture in a clear manner [12–15] Instead of an in-depth analysis and evaluation of various kinds of reasoning and fields concerned with it, here the aim is to describe what is new about the reasoning that the RSER system facilitates Systems such as the semantic web use the semantics that information is labelled with to reason for the best possible outcome, the RSER system reasons in a similar manner using the document structures directly Furthermore, the system communicates its reasoning to the user who can change the underlying logic through which the system has produced the EoI ranking This means that reasoning becomes a cooperative process between the system and the user, providing an additional layer of transparency.

6. Evaluation

The aim of this section is to evaluate the RSER search system, which — if developed further — could be used by investigators with no prior knowledge of the structure of the data to rank interesting entities The intention here is not to "solve" or "automate" the task of investigative retrieval, but to demonstrate the validity of the proposed methodology

6.1. Test collections

As discussed in Section 2, the data structures that InvR deals with are highly varied, often containing document types such as emails, legal agreements, spreadsheets, message chains etc In an ideal scenario the proposed system would be evaluated on a test collection that has been used in large scale investigation such as the Panama Papers, or Snowden files However, the raw data for these kinds of information is not openly available Furthermore, as the retrieval task for the proposed search system tackles is non-standard, the ground truth, i.e. the optimal ranking of EoIs, has to be defined by us In order to do this we must possess enough knowledge of the area in question to know what is a good ranking of the entities, meaning the information has to be from an area that the author is familiar with, or even better an area that most readers will be familiar with This is why we consider movie and Wikipedia data, rather than more complex topics covered by previous investigations There is a well-known benchmark data collection that relates to InvIR; the Enron email data set The following details the dataset and explains why we cannot use it.

The Enron email data collection is a collection of emails and other electronic communications from the Enron Corporation, a company that was involved in one of the biggest corporate scandals in American history [16] The collection consists of over 500k emails and other documents that were collected during the investigation of the company's fraudulent accounting practices The Enron email data collection has been widely studied and analysed by journalists and others interested in understanding the scandal and its aftermath Furthermore, it has been used by academics as a benchmark collection for various tasks such as classification, message threading, network analysis, topic modelling etc [17–21]. At first glace it would seem like an ideal test collection for the evaluation here However, there are three reasons why it is not suitable:

1. None of the existing benchmark versions of the Enron data give a ground truth that fits the InvR task described in this chapter.
2. The author does not possess enough domain knowledge to define base queries, seed entities, potential EoIs or the correct final rankings for the Enron email data.
3. The structure of the data (sender, receiver, subject, body, date) does not have the complexity required to infer "context" to the extent that our system requires.

Since the purpose of this evaluation is not to show that the proposed system "solves" InvIR, but to demonstrate that RSER is useful in an InvIR scenario both in terms of general performance and visualisation, there is no need to use data collections directly related to existing investigations For this reason we have chosen datasets that a non-expert readers are familiar with, thus making the evaluation more transparent The underlying datasets used are DBpedia and IMDB.

6.1.1. Example topic to be searched

Due to the complexity of the retrieval task, the topics formulated for testing the performance of the system are more complex than in traditional (ad-hoc) IR test collections Listing 1 shows the structure of a test topic.

For each topic we have the base query ("United States of America female crime thriller"), a seed entity (Uma Thurman), a list of interesting documents that a user would have checked (Kill Bill etc.) and a list of potential entities For each potential entity we have defined whether they match the information need, which in this instance corresponds to female actors in crime thrillers from the US Relevance is judged at 2 levels, 1 = relevant (female actors in movies that clearly fit the crime thriller genre and take place in the US), and 0 (not relevant) Altogether there are 15 topics for IMDB and 10 for DBpedia.

6.1.2. DBpedia dataset

The DBpedia test collection was first created by [22,23] There consists of 4.6 million entities There are 5 five document fields: {*names*, *related categories*, *similar entity names*, *entity name* and *attributes*} We have produced 15 topics that fit our investigative scenario Topics 9–14 are specifically designed to fit an investigative journalism scenario similar to that of the Panama Papers. The topics can be found in Appendix A and on GitHub.

6.1.3. IMDB dataset

The underlying data for the evaluation is an IMDB database collected by [24] The data consists of movie, actor and character data It has been cleaned and is stored in ElasticSearch instance and has the following fields: movie_id, plot, movie_name, movie_languages, movie_countries, movie_genres, actors, characters, actor_genders Altogether there are 15 topics which can be found in Appendix B The complete set of topics will be made available on github.

6.2. Configurations for testing

As a part of the evaluation we wish to test the performance of the system with various settings These settings are defined by four parameters which can also be chosen by the user on the interface The are as follows:

Retrieval Model: For the retrieval model we consider FSA(Field Score Aggregation)-BM25, ICFW-G-BM25, ICFW-GA-BM25, ICFW-LA-BM25.
Relevance: Three ways of defining r_i are considered.

1. Field-based BM25 retrieval scores.
2. Field weights assigned by the ICFW model.
3. A product of the two.

Similarity: For the similarity metric γ we consider manhattan distance, cosine distance and a combined metric where the two are multiplied.
k-cutoff: Finally, we try different values of k between 1 and 10. Fig. 4 shows system performance for different values of k The rest of analysis sets $k = 4$ as here we observe good performance for both data collections.

6.3. Experimental results

6.3.1. General performance

Table 5 shows the performance of RSER for different underlying retrieval models and similarity metrics at k = 4.

From Table 5 and Fig. 4 it is clear that there is significant variation in the performance of the system depending on how the underlying features are defined For this reason, it is important that the user interface offers options on these features for the user, as is done on the user interface in Fig. 3.

6.3.2. Is there value in using the field weights inferred by ICFW to define relevance structures?

As discussed earlier, ICFW was the stronger candidate compared to FSA-BM25 as the field weights are less noisy For FSA, field-based BM25 scores were used to estimate the relevance structure ($r_{i,\text{FSA}} = \text{RSV}_{\text{BM25}}(q, f_i, c)$ in Definition 4.1), whereas for the ICFW based models the ICFW field weights were used ($r_{i,\text{ICFW}} = w_{\text{ICFW}}(q, f_i, c)$ in Definition 4.1) We also experimented with a combination of ICFW-field weights and the BM25 score, where the two were multiplied to estimate $r_{i,\text{ICFW}-\text{BM25}}$: ($r_i = \text{RSV}_{\text{BM25}}(q, f_i, c) \times w_{\text{ICFW}}(q, f_i, c)$ in Definition 4.1).

Table 5 demonstrates the feasibility of using ICFW-based field weights for defining the relevance structures in RSER The results are relatively noisy due to the small number of evaluation topics per data-collection (especially the P@2 column) However, we can observe that there are no instances where the FSA-BM25-based model out performs the ICFW-based models It is worth noting that there are important

```
{
  "query_id": "3",
  "query": "United States of America female crime thriller",
  "seed_entity": "Uma Thurman",
  "documents_of_interest": [
    "UObKdYQB4U5l6a4-MCtu",
    "I-bLdYQB4U5l6a4-YEQo",
    "QebIdYQB4U5l6a4-0A4P",
    "X-bLdYQB4U5l6a4-DD1U",
    "buXGdYQB4U5l6a4-ueHa"
  ],
  "entities_of_interest": [
    "Lorraine Bracco",
    "Diane Keaton",
    "Jodie Foster",
    "Marlon Brando",
    "Robert Deniro",
    "Ray Liotta",
    "Joe Pesci"],
   "qrels": [
    [
      "3",
      "",
      "Lorraine Bracco",
      "2"
    ],
    ...
    [
      "3",
      "",
      "Robert Deniro",
      "0"
    ]]}
```

Listing 1: Example of a Topic for IMDB

Table 5
Experimentation results for RSER Overall the ICFW-LA model has performs the best, although due to the relatively small sample size in each data-collection significance cannot be inferred.

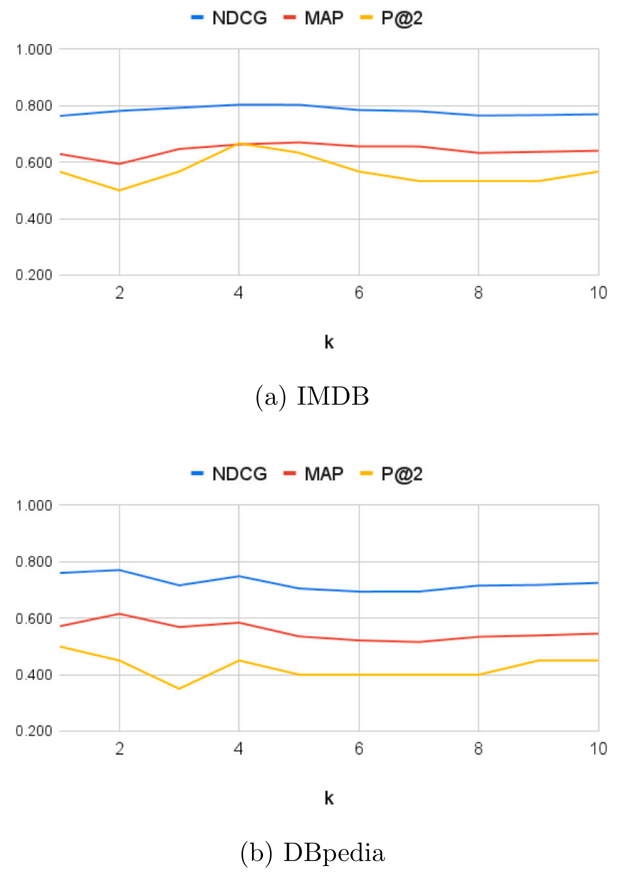
Relevance metric	rel. structure sim. metric	DBpedia			IMDB		
		ndcg	p2	map	ndcg	p2	map
FSA-BM25	cosine	0.766	0.533	0.544	0.750	0.560	0.567
	manhattan	0.759	0.667	0.553	0.716	0.567	0.540
	cos*man	0.761	0.633	0.560	0.674	0.600	0.443
ICFW-G	cosine	0.720	0.533	0.518	0.783	0.567	0.621
	manhattan	0.824	0.667	0.664	0.736	0.533	0.559
	cos*man	0.801	0.633	0.634	0.704	0.567	0.496
ICFW-GA	cosine	0.788	0.567	0.637	0.770	0.567	0.622
	manhattan	**0.844**	**0.700**	**0.729**	0.713	0.500	0.534
	cos*man	0.802	0.667	0.679	0.683	0.467	0.464
ICFW-LA	cosine	0.743	0.433	0.573	**0.803**	0.633	**0.670**
	manhattan	0.839	**0.700**	0.710	0.731	0.533	0.567
	cos*man	0.802	**0.700**	0.667	0.675	0.500	0.469
ICFW-G x BM25	cosine	0.768	0.467	0.605	0.660	0.589	0.415
	manhattan	0.787	0.633	0.623	0.689	0.533	0.497
	cos*man	0.819	**0.700**	0.651	0.689	**0.733**	0.405
ICFW-GA x BM25	cosine	0.760	0.533	0.580	0.662	0.700	0.381
	manhattan	0.756	0.500	0.606	0.663	0.333	0.460
	cos*man	0.743	0.467	0.575	0.662	0.700	0.381
ICFW-LA x BM25	cosine	0.753	0.600	0.581	0.656	0.633	0.429
	manhattan	0.762	0.533	0.612	0.708	0.533	0.522
	cos*man	0.755	0.500	0.584	0.643	0.700	0.376

difference between the two data collections: In general the benefits of using ICFW are greater for the IMDB dataset A likely reason for this is that the document structure for IMDB is much more complex than it is for DBpedia, with 9 fields for the former and 5 for the latter Furthermore, the fields are much more diverse for IMDB For example, the title of a Wikipedia page relates closely to its body, similar title and related titles, whereas for IMDB — apart from movie plot and title — the document fields are much more different semantically.

As a general trend we can say that ICFW-GA and ICFW-LA have the most robust performance across similarity metrics and datasets This makes sense intuitively as ICFW-LA considers the field level term metrics when calculating the lambda scaling parameter, whereas ICFW-G and ICFW-GA average over the fields See [4] for definitions.

6.3.3. *Which similarity model is the best one?*

The experimentation does not show that one similarity model is better than the others The results suggest that whether cosine similarity,



(a) IMDB

(b) DBpedia

Fig. 4. System performance for different k-values. k equals the cutoff of how many documents of interest are considered. See Section 6.2. $k = 4$ is chosen as a strong overall level for the rest of the evaluation.

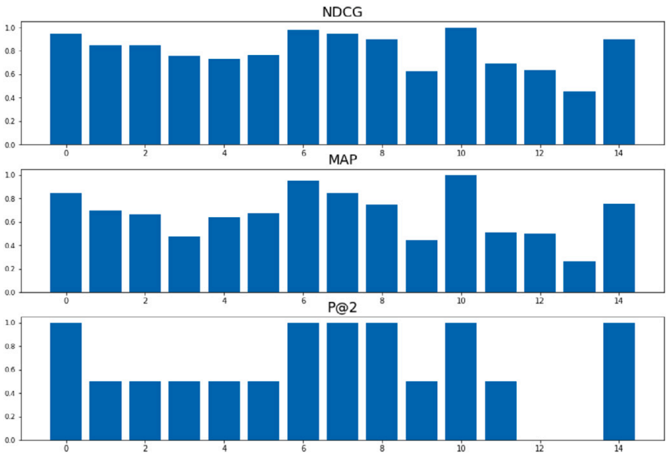


Fig. 5. RSER performance with ICFW-LA and cosine similarity on IMDB. Overall the performance is stable: For only two queries (12 and 13) the system is not able to rank one relevant entity in the top two.

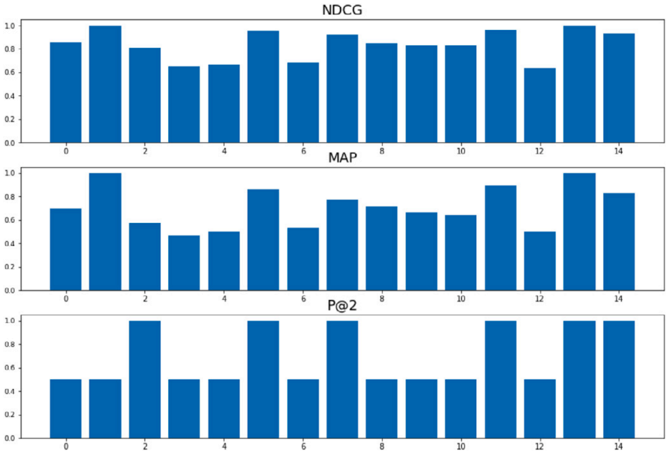


Fig. 6. Query-based Accuracies of RSER with ICFW-LA and manhattan similarity in DBpedia. Overall the performance is stable, with the system always ranking at least one relevant entity in the top two.

or manhattan distance is better, depends heavily on the data collection For DBpedia, manhattan distance would seem to produce better results when ICFW models are used and worse results if a combination of ICFW and BM25 is used Overall for DBpedia the best results are obtained by using ICFW-LA together with manhattan distance A possible reason for this is that due to the simpler document structure, the model needs to consider the degree of relevance for each field, as well as relative importance of each field For the IMDB dataset, cosine similarity does better than manhattan similarity This is likely to be because the relative importance of fields is more important than their degree.

To clarify this point, consider the query "United States President Haravard University" with the seed entity "Barack Obama" for DBpedia and the query "Actors in Italian mafia movies" with the seed entity "Al Pacino" For the former important documents would include Wikipedia articles such as "Barack Obama's timeline" for the latter movies such as "The Godfather" For the latter we would like to rank high entities where the query terms "Italian mafia movies" occur in the plot and/or description fields and the entity name (Al Pacino), occurs in the actor_names field If the entity name occurs in any other field, the context is automatically wrong, as we are only interested in actors, not characters for example What this means is that the relative importance of fields in terms of the relevance structure is of large importance, most likely more so than the actual degree of relevance for any individual field For DBpedia things are not as straight forward Since there are not as many fields, the system will find it more difficult to differentiate context based on whether a field is relevant or not, instead the degree of relevance will need to be considered For example, the Wikipedia timeline article should be about the EoI in question, but there is no field that lists the important entities in a document for example So the system would need to distinguish the occurrence of the entity name (Obama) from other query terms, such as president.

To summarise, with fewer fields the term level occurrences of query terms become more important that field level occurrences, which is likely to be the reason why for DBpedia manhattan similarity does better than cosine similarity and why for IMDB the opposite is true.

6.3.4. *Overall, what does the performance of RSER on the test collection tell us about its effectiveness in general?*

Figs. 6 and 5 show the query level accuracies for RSER with IMDB and DBpedia respectively We can see that for all three accuracy metrics, the performance of RSER is relatively steady across queries For only three of the queries does MAP drop below 0.45 and for NDCG below 0.75 Precision@2 is 0.5, or 1 for all queries apart from two (12 and 13), meaning the first two entities in the final ranking usually provide at least one good true positive entity. Queries 12 and 13 are "action film arnold Schwarzenegger male characters" and "german speaking movies set in Berlin during war times" respectively and the seed entities are "John Matrix" and "Good by Lenin" respectively Potential reasons for what could cause noise in the first two rankings is the combination of action and matrix for the first one, which are highly connected for other movies and the words german being used together with war Our stemming procedure means that German and Germany are equivalent, which is not ideal in this scenario However, these two issues aside the above suggests that RSER is indeed accurate enough to help investigators rank entities in terms of their context relative to a seed entity, albeit for two narrow test collections.

6.3.5. Discussion

The experimentation has demonstrated that the entity ranking aspect of RSER is able to rank movie and Wikipedia-based entities relatively well A user-based evaluation with investigation-related data would be required to unequivocally show that the system can be used in investigations Such a study is outside of the scope of this paper and is left for future research. However, the experimentation does suggest that the approach is valid and warrants further study.

7. Discussion

The purpose of this section is to frame the proposed methodology of structure-driven InvIR in the context of current research trends and IR research specifically Furthermore, investigative journalism is discussed as a potential application area.

7.1. Analytical retrieval methods and verification

In the context of generative AI and other black-box methods gaining more popularity in IR and elsewhere, this paper contributes to the wider objective of producing analytical and transparent methods which can complement various black-box systems in terms of verification, analysis and justification These aspects are important in many fields such as law enforcement, journalism and open-source investigations.

Black-box models can provide important insights and help users crawl through immense data collections quickly, finding answers and insights However, it is difficult to dissect exactly what the model has learned There exists a whole field of study for “explainable AI” and interpreting the results of black-box models See [25–27] for a summary and discussion What unites almost all of the approaches in the literature is that the interpreting and explaining of a model is done after the model has produced its outcomes The difference to the kind of transparency discussed in this paper is that the interpretation can be done at any stage The outcome of an analytical model is known in advance, as per our definition, meaning the degree of transparency and therefore interpretability is much greater than any form of “explainable AI” Analytical models therefore can be used to verify findings and to provide a layer of transparency needed for users and investigators to justify their findings in order for other people to trust them fully Leveraging document structures to give users a more comprehensive understanding of their findings can help them verify the outputs of black-box systems.

More concretely, consider an initial black-box method that is used to crawl a data-collection This could be a large language model-based search system, or a graph-based system leveraging an outside knowledge base for example Any findings a user makes from the collection are vulnerable to bias in the above described algorithms and/or in the underlying data they have been trained on In order to verify the findings in terms of the data collections itself, especially in terms of false negatives, transparent methods are required For example if a LLM-based search system provides a name that it thinks is of vital importance to the user, they can use a system such as RSER to check whether there are more entities that occur in a similar context and whether they are actually more interesting in terms of the specific data collection.

7.2. Relevance structures as extensions of traditional models

There exists a large body of research that extends well established analytical models such as the BM25 and Language Modelling in various ways On of the more prominent research directions is to bridge the semantic gap between the queries and relevant documents by adding a semantic component into the retrieval model Examples of such attempts include but are not limited to: [28–30].

This paper has taken a slightly different approach to extending existing analytical models, focusing on structure However, the semantic aspects of the documents are considered more extensively than by non-extended models, since the document structure does often communicate semantics as well For example, a name occurring in an actor field has a different semantic meaning than a name occurring in a character field Even though the methods discussed and proposed here do not consider these semantic meanings explicitly, they can differentiate between them This is something not done by the non-extended versions of BM25, or LM for example.

There are many methods that extend the BM25 and LM models in terms of structure, such as the BM25F and Mixture of Language Models (MLM) However, as discussed in Section 3.2 and in more detail by [3] unlike the methods used in this paper, these approaches do not satisfy all the SDR constraints and therefore their performance is limited [4] Furthermore, the methodology proposed in this paper take the consideration of document structures a step further: The structure is not only used to improve the performance of the retrieval system, but also to communicate to the user the context in which documents are relevant, also in a semantic sense This is accomplished by the concept of relevance structures.

7.3. InvIR and data-driven investigations

Due to data leaks, social media and the opening of various government databases, data-driven investigative methods have become available to a wider set of actors, including journalists Retrieval, search and discovery are vital aspects of these data-driven investigations (DDIs) As discussed in Section 2, InvIR and the methods proposed here emphasise transparency and reasoning, which is what makes them ideal for DDIs The Panama Papers project presents an informative example of a DDI The data the investigators trawled through was immense and largely unstructured, meaning the database had to be reconstructed before it could be effectively searched and reported on. This took a team of technical experts over a year to accomplish and involved a great deal of automation [31] It has been estimated that by 2021 different countries had recovered 1.3 billion dollars in tax revenue as a direct result of the Panama Papers [32] As journalists lack the authority of powerful institutions — unlike law enforcement for example — they have to be able to trust their findings and to back up the facts they are reporting to the rest of society Effectively, the burden of proof of their findings fully lies with the journalist and therefore they have to be able to trust and understand their the tools they use If the journalist cannot show the proof for their findings using underlying documents, the story cannot be published This is why they have to be able to understand in a transparent manner system they are using and they have to be able to reason with it In order to relate this paper to Panama Papers more concretely, evaluation topics 9–14 for the DBpedia dataset described in Section 6.1.2 and in Appendix A cover issues such as tax havens and tax evasion.

RSER accomplishes this by emphasising transparency and facilitating reasoning through interaction with the system After the system has ranked the list of EoIs using their relevance structure-based similarity to the seed entity, the user can easily follow the reasoning behind each ranking Furthermore, they can alter that reasoning to fit better with their own understanding of issues, thus giving the user a strong understanding of not only what entities are most interesting, but why as well.

8. Conclusion

This paper proposes a methodology for document structure-driven investigative retrieval (InvIR) The paper views InvIR as a sub-field of exploratory IR (ExIR), where more emphasis is given to transparency and reasoning. The first two sections focused on aspects of the proposed methodology that relate to document structures Firstly, existing research on analytical SDR and how it relates to the proposed methodology was discussed Secondly, the concept of relevance structures was introduced to describe the way in which the documents structure contributes to its relevance.

Having described the proposed methodology, the paper introduced a prototype retrieval system denoted Relevance Structure-based Entity Ranker (RSER), which is used to demonstrate and investigate the validity of the methodology An evaluation of the system demonstrated that the use of relevance structures together with analytical SDR models is able to provide the user with an additional layer of transparency and navigation Furthermore, system performance was good enough to warrant further study of the methodology and the kind of search system RSER represents.

The methodology proposed in this paper contributes to the wider area of computer science and IR by presenting analytical methods that are complimentary to black-box methods with regards to transparency and verification. Furthermore, the methodology has the potential to contribute to the application area of data-driven investigations (DDIs) directly through further study of systems such as RSER.

The structure-driven methodology for transparency-focused exploratory IR (InvIR) proposed and evaluated in this paper represents a promising and important new research direction in a world where transparency is instrumental in analysing the outcomes of various black-box systems.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

I would like to acknowledge the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1) for funding this research.

Appendix A. DBpedia topics

See Table A.6.

Appendix B. IMDB topics

See Table B.7.

Table A.6
Information needs, base queries and seed entities for DBpedia test collection.

ID	Information Need	Base Query	Seed
0	US presidents that went to Harvard University	United States President Harvard University	Barack Obama
1	United Nations general secretaries from Africa	united nations secretary general african	Kofi Annan
2	Books about the second world war in asia	second world war book pacific asia	Guadalcanal Diary
3	Authors of books about WW1	author book world war one	Barbara Tuchman
4	Countries with off shore oil rigs	countries with sea based oil reserves drilling platform	Norway
5	Countries in the Americas with oil	country america oil	Venezuela
6	Books about the Spanish civil war	book spanish civil war	For Whom the Bell Tolls
7	Organised crime figures in chicago	organised crime figure chicago mafia	Al Capone
8	Authors of books about the italian mafia	author book italian mafia	Mario Puzo
9	Mafia members involved in helping us military in WW2	mafia ww2 world war 2 war effort help new york docks	Luciano
10	Countries that are considered tax havens and are islands	tax haven island	Bermuda
11	People exposed in tax haven scandals and tax evasion	tax haven evasion scandal leak	Rami Makhlouf
12	Tax evasion leak whistle blowers	tax haven evasion leak whistle blower	Herve Falciani
13	News outlets involved in tax haven leaks	news organisation tax haven leak	Sud Deutsche
14	Countries involved in tax evasion	tax evasion haven	Seychelles

Table B.7
Information needs, base queries and seed entities for IMDB test collection.

ID	Information Need	Base Query	Seed
0	Actors that have appeared in westerns with Clint Eastwood	Clint Eastwood Western	Wallach
1	character names in italian mafia movies	italian mafia	Vito Corleone
2	actors names in italian mafia movies	italian mafia	Pacino
3	Female lead characters in movies about crime in the united states of america	United States of America female crime thriller	Uma Thurman
4	Movies with Harrison Ford that take place in United States with action	Harrison Ford United States action	Fugitive
5	characters in comedies with Jim Carrey about Christmas	comedy Jim Carrey Christmas	Grinch
6	actors in films with wizards and magic	wizard magic spells	Daniel Radcliffe
7	actors in films with wizards and magic	wizard magic spells	Hermione
8	characters in movies with Bruce Willis about Christmas	Christmas Bruce Willis	John McClane
9	movies with Bruce Willis and Samuel Jackson	Bruce Willis Samuel Jackson	Die Hard 3
10	marx brothers	black and white commedy	Harpo Marx
11	male actors in action films with Schwarzenegger	action film arnold Schwarzenegger	Dolph Lundgren
12	male characters in action films with Schwarzenegger	action film arnold Schwarzenegger	John Matrix
13	german speaking movies set in Berlin during war times	Berlin German Language War	Good bye Lenin
14	Characters that have appeared in westerns with Clint Eastwood	Clint Eastwood Western	Tuco

References

[1] G. Marchionini, Exploratory search: from finding to understanding, Commun. ACM 49 (4) (2006) 41–46.

[2] N. Fuhr, Salton award lecture information retrieval as engineering science, ACM SIGIR Forum 46 (2) (2012) 19–28.

[3] T. Ketola, T. Roelleke, Formal constraints for structured document retrieval, in: SIGIR, in: ICTIR, ACM, New York, NY, USA, 2022, pp. 121–126.

[4] T. Ketola, T. Roelleke, Automatic and analytical field weighting for structured document retrieval, in: Advances in Information Retrieval, ECIR '23, 2023, pp. 489–503.

[5] T. Ketola, T. Roelleke, BM25-FIC: Information content-based field weighting for BM25F, in: BIRDS@SIGIR, 2020, pp. 79–85.

[6] J. Kim, X. Xue, W.B. Croft, A probabilistic retrieval model for semistructured data, in: ECIR 2009, Springer, Berlin, Heidelberg, 2009, pp. 1–12.

[7] S. Robertson, H. Zaragoza, M. Taylor, Simple BM25 extension to multiple weighted fields, in: CIKM '04, ACM, Washington, D.C., USA, 2004, pp. 42–49.

[8] P. Ogilvie, J. Callan, Combining document representations for known-item search, in: SIGIR '03, ACM, New York, NY, USA, 2003, pp. 123–138.

[9] N. Zhiltsov, A. Kotov, F. Nikolaev, Fielded sequential dependence model for ad-hoc entity retrieval in the web of data, in: SIGIR '15, 2015, pp. 249–259.

[10] G. Osborne, B. Turnbull, J. Slay, Development of InfoVis software for digital forensics, in: 2012 IEEE 36th Annual Computer Software and Applications Conference Workshops, 2012, pp. 213–217.

[11] E. Higgins, We are Bellingcat: An Intelligence Agency for the People, Bloomsbury Publishing, 2021.

[12] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web, Sci. Am. 284 (5) (2001) 34–43.

[13] P. Ristoski, H. Paulheim, Semantic Web in data mining and knowledge discovery: A comprehensive survey, J. Web Semant. 36 (2016).

[14] M. Teichmann, M. Weber, M. Zöllner, R. Cipolla, R. Urtasun, MultiNet: Real-time joint semantic reasoning for autonomous driving, in: 2018 IEEE Intelligent Vehicles Symposium (IV), 2018, pp. 1013–1020.

[15] K. Li, Y. Zhang, K. Li, Y. Li, Y. Fu, Visual semantic reasoning for image-text matching, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4654–4662.

[16] B. Klimt, Y. Yang, The enron corpus: A new dataset for email classification research, in: Machine Learning: ECML 2004, in: Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2004, pp. 217–226.

[17] S. Alkhereyf, O. Rambow, Work hard, play hard: Email classification on the avocado and enron corpora, in: Proceedings of TextGraphs-11: The Workshop on Graph-Based Methods for Natural Language Processing, ACM, Vancouver, Canada, 2017, pp. 57–65.

[18] V. VanBuren, D. Villarreal, T.A. McMillen, A.L. Minnicks, Enron Dataset Research: E-mail Relevance Classification, Texas State University Faculty Publications-Computer Science, 2009.

[19] J. Shetty, J. Adibi, Discovering important nodes through graph entropy the case of Enron email database, in: LinkKDD '05, ACM, New York, NY, USA, 2005, pp. 74–81.

[20] R. Bekkerman, Automatic categorization of email into folders: Benchmark experiments on enron and SRI corpora, in: Computer Science Department Faculty Publication Series, 2004.

[21] A. McCallum, X. Wang, A. Corrada-Emmanuel, Topic and role discovery in social networks with experiments on enron and academic email, J. Artificial Intelligence Res. 30 (2007) 249–272.

[22] F. Hasibi, F. Nikolaev, C. Xiong, K. Balog, S. Bratsberg, A. Kotov, J. Callan, DBpedia-entity v2: A test collection for entity search, in: SIGIR '17, 2017, pp. 1265–1268.

[23] K. Balog, R. Neumayer, A test collection for entity search in DBpedia, in: SIGIR '13, ACM, New York, NY, USA, 2013, pp. 737–740.

[24] D. Bamman, B. O'Connor, N.A. Smith, Learning latent personas of film characters, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 352–361.

[25] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, XAI—Explainable artificial intelligence, Science Robotics 4 (37) (2019) eaay7120.

[26] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, J. Zhu, Explainable AI: A brief survey on history, research areas, approaches and challenges, in: Natural Language Processing and Chinese Computing, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 2019, pp. 563–574.

[27] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, R. Ranjan, Explainable AI (XAI): Core ideas, techniques, and solutions, ACM Comput. Surv. 55 (9) (2023) 194:1–194:33.

[28] M. Bahrani, T. Roelleke, FDCM: Towards balanced and generalizable concept-based models for effective medical ranking, in: CIKM '20, ACM, New York, NY, USA, 2020, pp. 1957–1960.

[29] H. Fang, C. Zhai, Semantic term matching in axiomatic approaches to information retrieval, in: SIGIR '06, ACM, New York, NY, USA, 2006, pp. 115–122.

[30] J. Kamps, M. Koolen, S. Geva, R. Schenkel, E. SanJuan, T. Bogers, From XML retrieval to semantic search and beyond, in: Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF, INEX, 2019, pp. 415–437.

[31] F. Obermaier, B. Obermayer, Panama Papers: Breaking the Story of how the Rich and Powerful Hide their Money, Oneworld Publications, 2016.

[32] L. Sisti, P. Biondani, E. Diaz-Struck, Counting the Panama Papers money: how we reached $1.24 billion, ICIJ (2019).