A formal model for data fusion

Mounia Lalmas

Department of Computer Science, Queen Mary University of London,
London E1 4NS, England, United Kingdom
mounia@dcs.qmul.ac.uk
http://www.dcs.qmul.ac.uk/

Abstract. In information retrieval, the data fusion problem is as follows: given two or more independent retrieved sets of ranked documents in response to the same query, how to merge the sets in order to present the user with the most effective ranking? We propose a formal model for data fusion that is based on the knowledge that can be derived from the retrieved documents. The model is based on evidential reasoning, a theory that formally expresses knowledge and the combination of knowledge. Knowledge characterising a ranked list of retrieved documents is symbolised. The combination of knowledge associated to the several retrieval results yields the characterisation of the merged result.

1 Introduction

The *information retrieval* problem is the quest to find the set of documents relevant to a query. The retrieved documents upon delivery to the user are ranked according to how they have been estimated relevant to the query. A new challenge in information retrieval has arisen with distributed document collections. Some documents may belong to several collections, whereas others belong to a single collection. Also, representation methods and retrieval strategies may vary from collection to collection. Therefore, each collection often produces a different set of documents and a different ranking of documents as a response to a query. The challenge is the following: "how to combine the ranked sets in order to present the user with the most effective ranking?". This is referred in information retrieval as the *data fusion* problem [15, 17, 1, 3, 10, 12, 14, 8, 16].

This paper proposes a formal model for data fusion based on *evidential reasoning* as developed in [11]. Our approach is based on a formalisation of the *knowledge* that can be derived from the retrieval results. By symbolising the knowledge describing a retrieved set of documents, the *combination of knowledge* yields the knowledge that characterises the merged set of documents. Evidential reasoning formally expresses knowledge and the combination of knowledge.

The paper is organised as follows. First, we provide some background and motivation of our work. Second, we express formally the knowledge describing a retrieval result. Third, we show how the combination of knowledge as defined in evidential reasoning allows the expression of the merged result. Finally, we conclude with some thoughts for future work.

2 Background and motivation

The data fusion process is divided into three steps: (1) collection selection, which corresponds to the identification of the document collection(s) most likely to contain relevant documents; (2) document selection, which corresponds to determining the number of documents to be retrieved from the selected collection(s); and (3) merging, which corresponds to the actual combination of the individual ranked lists produced by the multiple collections

Collection selection can be implemented by ranking collections, using for example trained queries [17]. This step however increases query time, so the method usually adopted is to treat equally all collections / search engines / strategies (e.g. [15]). The number of selected documents can be determined to be proportional to the quality of each collection [3, 17], although most approaches still retrieve an equal number of documents from each collection. Finally, the merging can be based on rank position [17, 18], document score [12], or information such as the title or abstract [16]. Similar approaches have been followed by metasearch engines (e.g. ProFusion [6, 7], SavvySearch [5], MetaCrawler [13], Fusion [15]).

Various types of information and heuristics are exploited to perform data fusion. It is however not yet possible to determine which types of information or heuristics, or their combination, consistently lead to effective data fusion. This is also because the available information is often not comparable. For instance, many search engines do not return score information, and when they do, these scores are not comparable since they are based on different retrieval strategies as well as statistics upon which these strategies are based.

This paper proposes a formal model for data fusion that can be considered as a meta-model. Our aim is to go beyond the specific approaches that have been developed and implemented, thus obtaining a general framework for representing the data fusion process. The model is based on a formalisation of the knowledge that can be derived from individual ranked lists of retrieved documents. The knowledge describes rank position, score, term appearing in the title and the abstract of a retrieved document, the quality of the a retrieval engine or retrieval strategy, the quality in terms of coverage of the collection itself, etc; the knowledge describes a retrieval result. We can then combine the knowledge describing each retrieval result to arrive at the description of the merged result.

The model proposed in this paper is not specific to a collection or a retrieval strategy. We want a general framework in which we can study the data fusion process. The development of a formal model is not new in information retrieval and has already been shown to lead to the generalisation of information retrieval models. As a result, it was possible to study these models, thus gaining insights about the information retrieval process (see for example [9]). Our ultimate aim is to determine which properties lead to effective data fusion, and which ones do not, thus obtaining a better understanding of the data fusion process. This work is a first step toward this aim, through the development of a formal framework.

3 Representing a single retrieval result

We start by formalising the knowledge that describes a retrieval result. Let a document collection be represented as a set of documents D. From a retrieved set of ranked documents, properties can be derived concerning the documents in D. The derived properties constitute *knowledge* held by an entity, an *agent*, observing the retrieved documents. For instance, from the ranking alone, the agent knows which documents have been estimated more relevant than others. If the titles of the retrieved documents are displayed, the agent may observe (and hence know), for example, that documents containing the term "wine" in their title are ranked higher than those containing the term "water".

In *evidential reasoning*, the knowledge of an agent is formally defined by an *epistemic operator* K. Evidential reasoning as developed by Ruspini [11] provides concepts that formalise the knowledge of the agent. We use these concepts to model the knowledge associated to a retrieval result: the properties describing documents, the documents themselves and the ranking.

3.1 Representing properties

Knowledge about a retrieval result consists of *known properties* of documents. The properties are formalised upon a *proposition space* and a *sentence space*.

Definition 1 *Let* $P = \{p_0, \ldots, p_n\}$ *be a* proposition space*, where the* p_i*s are propositions.*

The propositions formalises basic properties qualifying a retrieval result. Which properties are effective in describing a retrieval result is an open research question. They also depend on what is available to the agent: ranking alone [15], retrieval status values [2], document titles, parameters associated with the underlying retrieval algorithm, or past retrieval sessions [17]. The properties can include: estimated amplitude of relevance based on some normalisation of the retrieval status values or/and the rank of the retrieved documents [12], effectiveness measure such as precision and recall values (when the underlying retrieval algorithm is known), document content (terms in title, abstract, summary). In this paper, we assume that, for each retrieval result, the properties have been identified and are modelled as propositions of P.

Complex properties of a retrieval result are expressed as objective *sentences* defined upon the proposition space.

Definition 2 *The set of objective sentences is defined as follows: (i) any proposition* p_i *in* P *is an objective sentence; (ii) if* ϕ *and* ψ *are objective sentences, then so are* $\phi \vee \psi$*,* $\phi \wedge \psi$*,* $\neg\phi$*, and* $\phi \rightarrow \psi$*.*

Not all objective sentences (whether they symbolise basic or complex properties of the retrieval set) constitute knowledge. Even if a property is true, it is

only when it is known true (by the agent) then its corresponding objective sentence becomes knowledge. This is formally expressed by an *epistemic operator* associated with the agent observing the retrieved result.

Definition 3 *Given an objective sentence ϕ, the sentence* $\mathrm{K}\phi$ *represents that the properties symbolised by ϕ are* known *by the agent. The set of objective sentences and sentences of the form* $\mathrm{K}\phi$ *constitute a* sentence space *denoted S.*

We symbolise the known properties of a document $d \in D$ by a sentence of S: the conjunction of the propositions modelling the known properties of d. For example, let these propositions be p_3, p_4 and $\neg p_8 \vee p_{10}$ for $p_3, p_4, p_8, p_{10} \in P$. Then the sentence $p_3 \wedge p_4 \wedge (\neg p_8 \vee p_{10})$ constitutes knowledge: $\mathrm{K}(p_3 \wedge p_4 \wedge (\neg p_8 \vee p_{10}))$. We denote the function $property : D \mapsto S$ to assign such a sentence to each document of the collection. In our example, $property(d) \equiv p_3 \wedge p_4 \wedge (\neg p_8 \vee p_{10})$. In the (worst) case when nothing is known about d (for example d has not been retrieved), $property(d) \equiv \bot$, where $\bot$ represents the *false* proposition. The *true* proposition is denoted $\top$.

Basic and complex properties which can characterise a retrieval result (retrieved and non-retrieved documents) have been symbolised, as well as those that are known to characterise the retrieval result. The next step is to formally represent documents.

3.2 Representing documents

We use *possible worlds* to formalise documents.

Definition 4 *A possible world is an interpretation $w : S \mapsto \{t, f\}$ that satisfies the axioms of the modal logic system* S5 [4]*, where t and f denote the truth values* true *and* false*, respectively. The set of all possible worlds for S, called the* universe*, is denoted U.*

Given a proposition space P, there can be a maximum of $2^{|P|}$ possible worlds (e.g., one in which all the p_is are true, one in which $p_2, \ldots, p_n$ are true and $\neg p_1$ is true, etc.). This number can be smaller when, for example, two propositions symbolise properties that can never be used jointly to describe a document. For example, if the sentence $p_i \rightarrow p_j$ is true (e.g., $p_i =$"retrieved after rank i" and $j < i$), we cannot have p_i true and p_j false in the same world.

A document $d \in D$ is represented by the set of possible worlds in which the sentence $property(d)$ is true. This is formally defined by a function $world : D \mapsto U$.

Definition 5 *For $d \in D$, $world(d) = \{w \in U | w(property(d)) = t\}$. Furthermore, the sentence* $\mathrm{K}(property(d))$ *is true in all worlds in $world(d)$.*

Suppose that $P = \{a, b\}$ with four possible worlds given in Table 1 (e.g., $w_1(a) = w_1(b) = t$; $w_2(a) = t$ and $w_2(b) = f$; etc.). Let $D = \{d_1, d_2, d_3\}$ where $property(d_1) \equiv a \wedge b$, $property(d_2) \equiv a$ and $property(d_3) \equiv \bot$ (d_3 has not been

Table 1. Possibles worlds for $P = \{a, b\}$ and $U = \{w_1, \ldots, w_4\}$

Worlds in U	True sentences
w_1	$a, b, \mathrm{K}(a \wedge b), \mathrm{K}a$
w_2	$a, \neg b, \mathrm{K}a$
w_3	$\neg a, b$
w_4	$\neg a, \neg b$

retrieved). Then $world(d_1) = \{w_1\}$, $world(d_2) = \{w_1, w_2\}$ and $world(d_3) = \emptyset$ ($\bot$ is true in no world). In addition, $\mathrm{K}(a \wedge b)$ and $\mathrm{K}a$ are true in w_1, and w_1 and w_2, respectively. Note that and $\top$ and $\mathrm{K}\top$ are true in all worlds.

Additional knowledge about a document can be inferred. This is because some sentences can imply others ($a \wedge b \rightarrow a \vee b$ is true in all worlds). From the axioms of the modal logic system S5, if $\mathrm{K}(\phi_1 \rightarrow \phi_2)$ is true (i.e., the agent knows that properties symbolised by ϕ_1 implies properties symbolised by ϕ_2) then so is $\mathrm{K}\phi_1 \rightarrow \mathrm{K}\phi_2$. Hence if $\mathrm{K}(a \wedge b)$ is true in w_1 (where $w_1 \in world(d_1)$) then $\mathrm{K}(a \vee b)$ is also true in w_1, thus providing additional knowledge about document d_1.

Some of the sentences known true in a world $w \in world(d)$ play a particular role for representing the document $d \in D$.

Definition 6 *The sentence ϕ logically implies the sentence ψ, denoted $\phi \Rightarrow \psi$, iff if ϕ is true at a possible world w, then ψ is also true in that world w.*

Definition 7 *An objective sentence ϕ is said the* most specific sentence *in w if it is known in w and for every objective sentence $\psi \in S$, the sentence $\mathrm{K}\psi$ is true in w iff $\phi \Rightarrow \psi$.*

For any world $w \in U$, such a sentence always exists because it can always be constructed as the conjunction of all sentences ψ_i such that $\mathrm{K}\psi_i$ is true in w (see the proof in [11]). The most specific sentence of a world $w \in world(d)$ corresponds to the most specific knowledge associated to the document d *with respect to that world*. A document can have several most specific sentences.

In our previous example, we have $a \wedge b$ and a as the most specific sentences of w_1 and w_2, respectively, and $w_1, w_2 \in world(d_2)$; hence, d_2 has two most specific sentences, one with respect to w_1, and one with respect to w_2.

The set of most specific sentences of a document is symbolised by the function $mss : D \mapsto \wp(S)$. For $d \in D$, the set of most specific sentences for document d, $mss(d)$, can be derived from $world(d)$. A definition is required first.

Definition 8 *$e(\phi)$ denotes the epistemic set containing all possible worlds that have ϕ as their most specific sentence.*

Some epistemic sets may be empty. In our example, we have $e(b) = \emptyset$.

Theorem 1 *For $d \in D$, $mss(d) = \{\phi \in S | e(\phi) \subseteq world(d)\}$.*

Proof: For $d \in D$ and $\phi \in mss(d)$, there exists $w \in world(d)$ such that ϕ is the most specific sentence in w: $w \in e(\phi)$. For all $w' \in e(\phi)$, all the known sentences of w and w' are those implied by ϕ. Therefore, $w(\mathrm{K}\psi) = w'(\mathrm{K}\psi)$ for all $\psi \in S$ such that $\phi \Rightarrow \psi$. Therefore, $w' \in world(d)$ so we obtain $e(\phi) \subseteq world(d)$[1].

To prove the reverse, we show that for $\phi \in S$, if $e(\phi) \not\subseteq world(d)$ then $\phi \notin mss(d)$. For $\phi \in S$ such that $e(\phi) \not\subseteq world(d)$, there exists a world $w \in e(\phi)$ such that $w \notin world(d)$. For all other worlds $w' \in e(\phi)$, w and w' yield the same truth value for sentences of the form $\mathrm{K}\psi$. If w is not used to represent d, then so is w': $w' \notin world(d)$. Since $e(\phi)$ contains all worlds that have ϕ as their most specific sentence, and none of them belong to $world(d)$, then ϕ cannot be a most specific sentence for d; hence $\phi \notin mss(d)$. □

In our previous example, $mss(d_2) = \{a \wedge b, a\}$. We have also $e(a \wedge b) = \{w_1\}$ and $e(a) = \{w_2\}$. Both $e(a \wedge b) \subseteq world(d_2)$ and $e(a) \subseteq world(d_2)$.

We have modelled the documents of a retrieval result and their known properties. We can also formally reason about the properties. The next step is to represent the ranking itself.

3.3 Representing the ranking

So far we have used only the logical basis of evidential reasoning. This theory has a second basis, probability theory. A probability distribution is defined upon an algebra of U (see [11] for details), upon which a mass function is constructed to quantify the uncertainty of sentences in S:

Definition 9 *A* mass function *is a mapping* $m : S \mapsto [0, 1]$ *such that:*

$$m(\bot) = 0 \quad \textit{and} \quad \sum_{\phi \in S} m(\phi) = 1$$

The mass function is not a probability distribution, and knowing the details of its construction is not necessary to the understanding of this paper and hence is omitted. What should be retained about m is that for $\phi \in S$:

$$m(\phi) = \begin{cases} > 0 & \text{if } e(\phi) \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

Only sentences that are most specific (with respect to some worlds) have a non-null mass value.

We use m to capture the impact of properties in ranking documents. The basic idea is that a sentence describing document d, which *should* be ranked higher than document d', has a higher mass value to that of a sentence describing document d'. Let $\phi, \phi' \in S$ such that $e(\phi) \neq \emptyset$ and $e(\phi') \neq \emptyset$. This is formally expressed as follows:

$$\begin{cases} m(\phi) \geq m(\phi') & \text{if } \phi \text{ qualify documents that should be retrieved (at least)} \\ & \text{before those qualified by } \phi', \\ m(\phi) < m(\phi') & \text{otherwise.} \end{cases}$$

[1] In fact, for any sentence $\phi \in S$, either $e(\phi) \subseteq world(d)$ or $e(\phi) \cap world(d) = \emptyset$.

The estimation of the mass function m requires a study of the document collection and the retrieval set from (whatever) evidence is available to the agent. The worst case is when only ranking information is available. The only properties that can be extrapolated concern the ranks themselves: a proposition p_i expresses the property that a document is ranked ith. Let a document d be ranked ith. We set $property(d) \equiv \phi_i$ where $\phi_i \equiv \neg p_1 \wedge \ldots \wedge \neg p_{i-1} \wedge p_i \wedge \neg p_{i+1} \wedge \ldots \wedge \neg p_l$ where l is the number of retrieved documents; ϕ_i means that the document has been retrieved at rank i and no other rank. If N is the number of documents in the collection, then we have N worlds w_i where $w_i(p_i) = t$ and $w_i(p_j) = f$, for $i \neq j$ (we cannot have p_i and p_j for $i \neq j$ both true in a world). Furthermore, for a document d retrieved at rank i, $world(d) = \{w_i\}$ and ϕ_i is the most specific sentence in w_i. $m(\phi_i)$ can be estimated as follows:

$$m(\phi_i) = \frac{l - (i-1)}{1 + \cdots + l}$$

which leads to the wanted inequality $m(\phi_i) > m(\phi_{i+1})$ since ϕ_i qualifies documents ranked before those qualified by ϕ_{i+1}. $\perp$ characterises the non-retrieved documents and by definition $m(\perp) = 0$.

The rank of a document is expressed in terms of a belief function defined upon the mass function.

Definition 10 *A belief function $Bel : S \mapsto [0,1]$ is defined as follows:*

$$Bel(\phi) = \sum_{\psi \Rightarrow \phi} m(\psi)$$

Definition 11 *The rank of a document is given by the function $R : D \mapsto [0,1]$. For $d \in D$[2], $R(d) = Bel(property(d))$.*

Documents are ranked in decreasing order of the value $R(.)$. If $R(d) = 0$ then document $d \in D$ is not retrieved.

Theorem 2 *For $d \in D$, $R(d) = \sum_{\phi \in mss(d)} m(\phi)$.*

To prove this theorem let us prove first the following lemma.

Lemma 1 *Let $d \in D$. For all $\phi \in mss(d)$, $\phi \Rightarrow property(d)$.*

Proof: Let $\phi \in mss(d)$. Let $w \in U$ such that ϕ is true in w. We have two cases: (1) $w \in e(\phi)$ so from Theorem 1 $w \in world(d)$, and hence K($property(d)$) is true in w. (2) $w \notin e(\phi)$, then there exists $\phi' \in mss(d)$ such that $\phi' \Rightarrow \phi$. We have also $e(\phi') \subseteq world(d)$ from Theorem 1, so for the same reason as above, K($property(d)$) is true in w. Therefore, for any world where Kϕ is true, K($property(d)$) is true, thus $\phi \Rightarrow property(d)$. $\Box$

[2] Ranks are usually integer. In our case they are real numbers, but this is not a problem because what is important is to have an ordering.

We prove now Theorem 2.

Proof: Let $d \in D$. From the definition of Bel, $R(d) = Bel(property(d)) = \sum_{\phi \Rightarrow property(d)} m(\phi)$. With Lemma 1, for $\phi \in mss(d), \phi \Rightarrow property(d)$. All sentences in $mss(d)$ are most specific, so $e(\phi) \neq \emptyset$, hence $m(\phi) > 0$. Each such $m(\phi)$ contributes to the value of $Bel(property(d))$:

$$Bel(property(d)) \geq \sum_{\phi \in mss(d)} m(\phi)$$

A most specific sentence $\phi \notin mss(d)$ cannot implies $property(d)$, and hence does not contribute toward the value of $Bel(property(d))$. All none most specific sentences, whether they imply or not $property(d)$ have null mass value. Therefore, we obtain the wanted equality. □

Theorem 2 shows that in our fusion approach, the rank of a document can be based on those most specific sentences that imply the document property sentence, when an appropriate mass function m is constructed to capture the impact of the most specific sentences in ranking documents. The mass function is built upon the knowledge (evidence) observable from a retrieval result. This complete the formal representation of a single retrieval result.

3.4 Working example

In the remaining of this paper, we will use a working example. Suppose that $D_1 = \{d_1, \ldots, d_4\}$ and $D_2 = \{d_1, \ldots, d_5\}$ (note that d_5 is not a document of collection D_1). We assume the following sets of worlds $U_1 = \{w_1, \ldots, w_6\}$ and $U_2 = \{w'_1, \ldots, w'_5\}$. Let the rankings from collections D_1 and D_2 be referred to as $\mathcal{R}_1$ and $\mathcal{R}_2$, respectively. The ranked documents, the worlds representing them, and the set of most specific sentences are shown in Table 2. In the example, for simplicity, each document has a unique most specific sentence. Therefore, for $d \in D_i$, $property_i(d)$ is a most specific sentence and the only most specific sentence in $mss_i(d)$. Note that the documents d_1 and d_4 have not been retrieved in ranking $\mathcal{R}_2$ ($property(d_1) \equiv property(d_4) \equiv \bot$).

Table 2. Ranked documents, the associated worlds and most specific sentences

Ranking $\mathcal{R}_1$			Ranking $\mathcal{R}_2$		
Documents	$world_1(d)$	$mss_1(d)$	Documents	$world_2(d)$	$mss_2(d)$
d_1	$\{w_1, w_2\}$	$\{\phi_1\}$	d_3	$\{w'_1\}$	$\{\phi'_1\}$
d_2	$\{w_1, w_2\}$	$\{\phi_1\}$	d_5	$\{w'_2, w'_3\}$	$\{\phi'_2\}$
d_3	$\{w_3\}$	$\{\phi_2\}$	d_2	$\{w'_2, w'_3\}$	$\{\phi'_2\}$
d_4	$\{w_4\}$	$\{\phi_3\}$			$\top$

The values of the mass functions m_1 and m_2 associated with $\mathcal{R}_1$ and $\mathcal{R}_2$ are given in Table 3. For instance, for ranking $\mathcal{R}_1$, the sentence ϕ_1 characterises

Table 3. Representation of the ranking

Ranking $\mathcal{R}_1$			Ranking $\mathcal{R}_2$		
Sentences	$e_1(.)$	$m_1(.)$	Sentences	$e_2(.)$	$m_2(.)$
ϕ_1	$\{w_1, w_2\}$	0.4	ϕ'_1	$\{w'_1\}$	0.5
ϕ_2	$\{w_3\}$	0.3	ϕ'_2	$\{w'_2, w'_3\}$	0.3
ϕ_3	$\{w_4\}$	0.2	$\top$	$\{w'_4, w'_5\}$	0.2
$\top$	$\{w_5, w_6\}$	0.1			

documents that should be ranked before those characterised by the sentence ϕ_2. From the above assumption (unique most specific sentence) and Theorem 2, for $d \in D_i$, $R_i(d) = Bel_i(property_i(d)) = m_i(property_i(d))$, which as required, produces the rankings shown in Table 2.

We discuss the use of $\top$ for instance for ranking $\mathcal{R}_1$. There is no knowledge with respect to properties covered by worlds w_5 and w_6. Since $\top$ is true in all worlds, then $\top$ is the only knowledge associated to these worlds, and then constitute the most specific sentence for both worlds.

For collection D_i, the knowledge associated to the ranked set of documents $\mathcal{R}_i$ is formalised with the epistemic operator K_i. The set of documents in the merged ranking is $D_1 \cup D_2$. The merging of two rankings can be viewed as a combination of knowledge. The next section describes the fusion of the two rankings.

4 Representing the merged retrieval results

Merging rankings in a model for data fusion based on evidential reasoning is formalised in terms of a *combination of evidence*. The latter is formalised by an epistemic operator K, where the knowledge of the *combined agent* K is defined in terms of the knowledge of the individual agents K_1 and K_2. As for the retrieval result of a single collection, the merged ranking has properties, documents, and a ranking which are determined upon those of K_1 and K_2.

4.1 Representing properties of the merged result

Let P_1 and P_2 be the proposition spaces associated with rankings $\mathcal{R}_1$ and $\mathcal{R}_2$. Let S_1 and S_2 be the respective sentence spaces.

Definition 12 *The sentence space of the combined ranking, denoted $S_\otimes$, is defined by the axioms: (i) if p is proposition of P_i, then it is a sentence of $S_\otimes$; (ii) axioms of logic defining well-formed sentences.*

(i) means that the propositions modelling properties in the merged result are those modelling properties in the individual retrieval results. (ii) defines sentences symbolising complex properties in the merged ranking.

A document in the combined ranking is also characterised by a sentence describing its known properties in the merged ranking. The sentence can be viewed as the *disjunction* of the two sentences describing the document in $\mathcal{R}_1$ and $\mathcal{R}_2$, respectively. First, we extend the definition of $property_i$ for $i = 1, 2$.

Definition 13 *Let* $property_i' : D_1 \cup D_2 \mapsto S_i$ *the extension of* $property_i$. *For* $d \in D_1 \cup D_2$*:*

$$property_i'(d) \equiv \begin{cases} property_i(d) & \text{if } d \in D_i, \\ \bot & \text{otherwise} \end{cases}$$

The sentence $\bot$ characterises a document in D_1 that is not in D_2 and vice versa. The sentence characterising a document in the merged ranking can now be defined.

Definition 14 *Let* $property_\otimes : D_1 \cup D_2 \mapsto S_\otimes$. *For* $d \in D_1 \cup D_2$, $property_\otimes(d) \equiv property_1'(d) \vee property_2'(d)$.

In our working example, we have $property_\otimes(d_1) \equiv \phi_1 \vee \bot \equiv \phi_1$ (d_1 is not retrieved in $\mathcal{R}_2$), $property_\otimes(d_2) \equiv \phi_1 \vee \phi_2'$, $property_\otimes(d_3) \equiv \phi_2 \vee \phi_1'$, $property_\otimes(d_4) \equiv \phi_3 \vee \bot \equiv \phi_3$ (d_4 is not retrieved in $\mathcal{R}_2$), and $property_\otimes(d_5) \equiv \bot \vee \phi_2' \equiv \phi_2'$ (d_5 is not in D_1). For a document $d \in D_1 \cup D_2$ that is retrieved in neither collection, $property_\otimes(d) \equiv \bot$.

4.2 Representing documents in the merged result

Let U_1 and U_2 be the universes associated with rankings $\mathcal{R}_1$ and $\mathcal{R}_2$. Documents in the combined ranking are also represented by a universe.

Definition 15 *The universe of the combined ranking is denoted* $U_\otimes$. *A possible world in* $U_\otimes$ *is a mapping* $w : S_\otimes \mapsto \{t, f\}$ *that satisfies the following axioms: (i)* w *satisfies the axiom of modal logic* S5*; (ii) if* ϕ *is a sentence of* $S_\otimes$, *then* $\mathrm{K}\phi$ *is true in* w *iff there exists* ϕ_1 *and* ϕ_2 *in* S_1 *and* S_2, *respectively, such that* $\mathrm{K}\phi_i$ *is true in* w *and* $\phi_1 \wedge \phi_2 \Rightarrow \phi$.

(ii) means that the knowledge about the combined ranking comes from the conjunction of sentences in S_1 and S_2 modelling properties in the two rankings, and any sentence that can be derived from the conjunction. For a document to be characterised by a sentence ϕ in the combined ranking, the document must be characterised by two sentences ϕ_1 and ϕ_2 in the two individual rankings (the combined agent K must know this fact), respectively, such that $\phi_1 \wedge \phi_2$ logically implies ϕ.

The worlds in $U_\otimes$ are constructed upon the worlds in U_1 and U_2 by way of a *cartesian projection*.

Definition 16 *The* cartesian projection *is the mapping* $\Pi : U_\otimes \mapsto U_1 \times U_2$ *where* $\Pi(W) = (w_1, w_2)$ *and* w_i *is the unique possible world in* U_i *such that for a sentence* $\phi \in S_i$*: (i)* ϕ *is true in* w_i *iff* ϕ *is true in* W*; (ii)* $\mathrm{K}_i\phi$ *is true in* w_i *iff* $\mathrm{K}\phi$ *is true in* W.

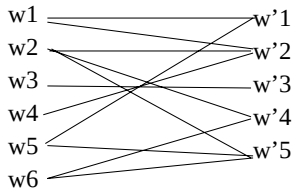

Fig. 1. Compatible worlds between U_1 and U_2

Table 4. Possible worlds in the combined ranking

$U_\otimes$	$\Pi(W_i)$	Kϕ	$U_\otimes$	$\Pi(W_i)$	Kϕ
W_1	(w_1, w'_1)	ϕ_1, ϕ'_1	W_2	(w_1, w'_2)	ϕ_1, ϕ'_2
W_3	(w_2, w'_2)	ϕ_1, ϕ'_2	W_4	(w_2, w'_4)	ϕ_1
W_5	(w_2, w'_5)	ϕ_1	W_6	(w_3, w'_3)	ϕ_2, ϕ'_2
W_7	(w_4, w'_2)	ϕ_3, ϕ'_2	W_8	(w_5, w'_1)	ϕ'_1
W_9	(w_5, w'_5)	$\top$	W_{10}	(w_6, w'_4)	$\top$
W_{11}	(w_6, w'_5)	$\top$			

(i) and (ii) means that the truth assignments to sentences of S_1 and S_2 in $U_\otimes$ worlds must respect those in U_1 worlds and U_2 worlds, respectively. From (ii), the maximum number of $U_\otimes$ worlds that can be created is $|P_1| \times |P_2|$. However, as for the number of possible worlds in the individual case, the number can be smaller since not all U_1 worlds are compatible with all U_2 worlds (e.g., one U_1 world in which p_i is true and one U_2 world in which p_i is false).

For our working example, we assume the compatibility between U_1 worlds and U_2 worlds given in Figure 1. w_1 is compatible with w'_1 and w'_2; w_2 is compatible with w'_2, w'_4 and w'_5; and so forth. The application of definitions 15 and 16 to our working example is shown in Table 4. Take $\Pi(W_1) = (w_1, w'_1)$. $\mathrm{K}_1\phi_1$ and $\mathrm{K}_2\phi'_1$ are true in w_1 and w'_1, respectively, so Kϕ_1 and Kϕ'_1 are true in W_1. Take now $\Pi(W_8) = (w_5, w'_1)$. $\mathrm{K}_1\top$ and $\mathrm{K}_2\phi'_1$ are true in w_5 and w'_1, respectively, so Kϕ'_1 is true in W_8. Note that, although not shown in Table 4, K$\top$ is true in all worlds.

Worlds in U_i are related to documents in D_i by way of the function $world_i$. Similarly, worlds in $U_\otimes$ are related to documents in $D_1 \cup D_2$ by way of a function $world_\otimes : (D_1 \cup D_2) \mapsto U_\otimes$. For $d \in D_1 \cup D_2$ and $W \in U_\otimes$, $W \in world_\otimes(d)$ iff K($property(d)$) is true W.

In our example, we obtain: $world_\otimes(d_1) = \{W_1, \ldots, W_5\}$, $world_\otimes(d_2) = \{W_1, \ldots, W_7\}$, $world_\otimes(d_3) = \{W_1, W_6, W_8\}$, $world_\otimes(d_4) = \{W_7\}$, and $world_\otimes(d_5) = \{W_2, W_3, W_6, W_7\}$.

Now that worlds composing $U_\otimes$ have been constructed, the most specific sentences for the combined ranking can be determined, as well as the corresponding epistemic sets $e_\otimes : S_\otimes \mapsto \wp(U_\otimes)$. Table 5 shows the sentence ϕ in $S_\otimes$ such that

Table 5. Most specific sentences and epistemic sets for the merged rank

ϕ	$e(\phi)$
$\phi_1 \wedge \phi_1'$	$\{W_1\}$
$\phi_1 \wedge \phi_2'$	$\{W_2, W_3\}$
ϕ_1	$\{W_4, W_5\}$
$\phi_2 \wedge \phi_2'$	$\{W_6\}$
$\phi_3 \wedge \phi_2'$	$\{W_7\}$
ϕ_1'	$\{W_8\}$
$\top$	$\{W_9, W_{10}, W_{11}\}$

$e_\otimes(\phi) \neq \emptyset$. For example, for worlds W_2 and W_3, the most specific sentence is the conjunction of ϕ_1 and ϕ_1'. Thus $e_\otimes(\phi_1 \wedge \phi_1') = \{W_2, W_3\}$.

A document in $d \in D_i$ has an associated set of most specific sentences $mss_i(d)$. Similarly a document $d \in D_1 \cup D_2$ has a set of most specific sentences. Let $mss_\otimes : D_1 \cup D_2 \mapsto \wp(S_\otimes)$ symbolising the set of most specific sentences for document in the combined ranking. For $d \in D_1 \cup D_2$, $mss_\otimes(d) = \{\phi \in S | e_\otimes(\phi) \subseteq world_\otimes(d)\}$.

4.3 Representing the merged ranking

As for single retrieval result, the combined ranking is defined in terms of a belief function defined upon a mass function. The mass function for the combined ranking is defined in terms of those of the individual rankings. We give first two definitions.

Definition 17 *Two sentences ψ and ϕ are* logically equivalent*, written $\psi \Leftrightarrow \phi$ iff $\psi \Rightarrow \phi$ and $\phi \Rightarrow \psi$.*

Definition 18 *The function $\Gamma : S_\otimes \mapsto \wp(S_1 \times S_2)$ maps every sentence ϕ in $S_\otimes$ to a subset of sentence pairs (ϕ_1, ϕ_2) with $\phi_i \in S_i$, $i = 1, 2$, such that $\phi_1 \wedge \phi_2 \Leftrightarrow \phi$ in $U_\otimes$.*

Γ relates sentences of $S_\otimes$ to those of S_1 and S_2.

In this work, we have assumed that the two retrieved sets are independent, which is the most common scenario in data fusion. With this assumption, the calculation of the mass function for the combined ranking is straightforward. However, it should be noted that the evidential reasoning as developed by Ruspini [11] allows for the dependent case to be taken into account.

Definition 19 *Let $m_\otimes : S_\otimes \mapsto [0, 1]$ be the mass function associated with the combined ranking. Let $m_1 : S_1 \mapsto [0, 1]$ and $m_2 : S_2 \mapsto [0, 1]$ be the mass functions associated with agents* K_1 *and* K_2*, respectively. For $\phi \in S_\otimes$:*

$$m_\otimes(\phi) = \mathcal{K} \sum_{(\phi_1, \phi_2) \in \Gamma(\phi)} m_1(\phi_1) \times m_2(\phi_2)$$

where $\mathcal{K} = \sum_{(\phi_1,\phi_2)\in S_1\times S_2} m_1(\phi_1) \times m_2(\phi_2)$*, ensuring that* $m_\otimes$ *is a mass function.*

The next theorem ensures that $m_\otimes$ is a mass function which as for m_1 and m_2 assigns non-null value only to most specific sentences in $S_\otimes$.

Theorem 3 *Let* $\phi \in S_\otimes$*. We have* $m_\otimes(\phi) > 0$ *if* $e_\otimes(\phi) \neq \emptyset$*.*

To prove Theorem 3, we prove first the following two lemmas.

Lemma 2 *Let* $\phi \in S_\otimes$ *and* $(\phi_1, \phi_2) \in \Gamma(\phi)$*. If* $e_\otimes(\phi) \neq \emptyset$*, then* $e_1(\phi_1) \neq \emptyset$ *and* $e_2(\phi_2) \neq \emptyset$*.*

Proof: Let $\phi \in S_\otimes$ such that $e_\otimes(\phi) \neq \emptyset$. Let $(\phi_1, \phi_2) \in \Gamma(\phi)$: $\phi_1 \wedge \phi_2 \Leftrightarrow \phi$. Suppose that ϕ_1 and/or ϕ_2 are not most specific sentences in S_1 and S_2 ($e_1(\phi_1) = \emptyset$ and/or $e_2(\phi_2) = \emptyset$). Then for all worlds in U_1 and/or all worlds in U_2, there exit sentences $\psi_1 \in S_1$ and/or $\psi_2 \in S_2$ such that $\psi_1 \Rightarrow \phi_1$ and/or $\psi_2 \Rightarrow \phi_2$. For all worlds $W \in U_\otimes$ such that ϕ is true in W, from Definition 15, ϕ_1 and ϕ_2 are true in W and from Definition 16, ψ_1 and/or ψ_2 are true in W. Therefore, $\psi_1 \wedge \phi_2 \Rightarrow \phi_1 \wedge \phi_2$, $\phi_1 \wedge \psi_2 \Rightarrow \phi_1 \wedge \phi_2$, and/or $\psi_1 \wedge \psi_2 \Rightarrow \phi_1 \wedge \phi_2$. From the relation between ϕ_1, ϕ_2 and ϕ, any of the latter means that there always exist sentences ψ_1 and/or ψ_2 such that $\psi_1 \wedge \phi_2 \Rightarrow \phi$, $\phi_1 \wedge \psi_2 \Rightarrow \phi$, and/or $\psi_1 \wedge \psi_2 \Rightarrow \phi$. That is, ϕ is not a most specific sentence, which contradicts the hypothesis $e_\otimes(\phi) \neq \emptyset$. Therefore, $e_1(\phi_1) \neq \emptyset$ and $e_2(\phi_2) \neq \emptyset$. □

Lemma 3 *Let* $\phi \in S_\otimes$*. If* $e_\otimes(\phi) \neq \emptyset$ *then* $\Gamma(\phi) \neq \emptyset$*.*

Proof: Let $W \in U_\otimes$ and $\phi \in S_\otimes$. From Definition 15, for Kϕ to be true in W, we have two sentences ϕ_1 and ϕ_2 in S_1 and S_2, respectively, such that $\phi_1 \wedge \phi_2 \Rightarrow \phi$. Furthermore, K$\phi_1$ and Kϕ_2 are true in world W. However, ϕ is the most specific world in W, then it must be the case that $\phi \Rightarrow \phi_1 \wedge \phi_2$. Therefore, the set $\Gamma(\phi)$ cannot be empty if ϕ is a most specific sentence. □

We prove now Theorem 3.

Proof: Let $e_\otimes(\phi) \neq \emptyset$. From Lemma 3, $\Gamma(\phi) \neq \emptyset$. Let $(\phi_1, \phi_2) \in \Gamma(\phi)$. From Lemma 2, ϕ_1 and ϕ_2 are most specific sentences with respect to U_1 and U_2, respectively: $m_1(\phi_1) \neq 0$ and $m_2(\phi_2) \neq 0$. From Definition 19, $m_1(\phi_1) \times m_2(\phi_2)$ contributes toward the value of $m_\otimes(\phi)$ which is hence non-null: $m_\otimes(\phi) > 0$. □

The calculation of $m_\otimes$ for our working example is shown in Table 6. $\Gamma(\phi)$ contains only one pair (ϕ_1, ϕ_2). The normalising constant is $\mathcal{K} = 0.62$. The results in Table 6 show that, for instance, in the combined ranking, documents characterised by $\phi_1 \wedge \phi_1'$ should be ranked before those characterised by $\phi_1 \wedge \phi_2'$. This result is plausible since from Table 3, we see that for ranking $\mathcal{R}_2$, ϕ_1' characterises documents that should be ranked before those characterised by ϕ_2'.

The ranking of a document d is defined as for the single retrieval case in terms of a belief function defined upon $m_\otimes$ as $R(d) = Bel_\otimes(property_\otimes(d))$ where $Bel_\otimes$ is the belief function associated with $m_\otimes$. Applied to our working example, the belief values are shown in Table 7 (the sentences $mss_\otimes(d)$ are shown for each

Table 6. Mass function calculation

ϕ	ϕ_1	ϕ_2	$m_1(\phi_1) \times m_2(\phi_2)$	$m_\otimes(\phi)$
$\phi_1 \wedge \phi_1'$	ϕ_1	ϕ_1'	0.4×0.5	0.32
$\phi_1 \wedge \phi_2'$	ϕ_1	ϕ_2'	0.4×0.3	0.19
ϕ_1	ϕ_1	$\top$	0.4×0.2	0.13
$\phi_2 \wedge \phi_2'$	ϕ_2	ϕ_2'	0.3×0.3	0.15
$\phi_3 \wedge \phi_2'$	ϕ_3	ϕ_2'	0.2×0.3	0.096
ϕ_1'	$\top$	ϕ_1'	0.1×0.5	0.08
$\top$	$\top$	$\top$	0.1×0.2	0.032

Table 7. Belief function values for the merged set

d	$mss_\otimes(d)$	R(d)
d_1	$\{\phi_1 \wedge \phi_1', \phi_1 \wedge \phi_2', \phi_1\}$	0.64
d_2	$\{\phi_1 \wedge \phi_1', \phi_1 \wedge \phi_2', \phi_1, \phi_2 \wedge \phi_2', \phi_3 \wedge \phi_2'\}$	0.89
d_3	$\{\phi_1 \wedge \phi_1', \phi_1', \phi_2 \wedge \phi_2'\}$	0.55
d_4	$\{\phi_3 \wedge \phi_2')\}$	0.096
d_5	$\{\phi_1 \wedge \phi_2', \phi_2 \wedge \phi_2', \phi_3 \wedge \phi_2'\}$	0.44

document d). The combined ranking is: d_2, d_1, d_3, d_5 then d_4. From the rankings $\mathcal{R}_1$ and $\mathcal{R}_2$, the possible combination of worlds from U_1 and U_2, and the mass functions m_1 and m_2, the combination of evidence as developed in evidential reasoning yields a coherent and intuitive combined ranking.

5 Conclusion

In this work, we have developed a *formal model* for data fusion based on evidential reasoning. The model considers the knowledge describing a retrieval result. The combination of knowledge determines the fusion of the retrieval results.

Our next step is the implementation of the model in order to study which properties of retrieved results lead to effective data fusion. The challenges will be (1) to map information and heuristics used so far in data fusion to properties, and (2) to determine appropriate mass functions that give higher values to properties that retrieve documents at higher rank. If the properties and the mass function are provided for each retrieval set, the combined ranking is a direct application of the proposed model.

References

1. Bartell, B., Cottrell, G., and Belew, R. Automatic combination of multiple ranked retrieval systems. In *Proc. 17th ACM SIGIR Conference* (Dublin,Ireland, 1994), pp. 172–181.
2. Baumgarten, C. A probabilistic model for distributed information retrieval. In *Proc. 21th ACM SIGIR Conference* (Philadelphia, USA, 1997), pp. 258–266.

3. Callan, J., Zhihong, L., and Croft, W. Searching distributed collection with inference networks. In *Proc. 18th ACM SIGIR Conference* (Seattle, USA, 1995), pp. 21–28.
4. Chellas, B. F. *Modal Logic: An introduction*. Cambridge University Press, 1980.
5. Dreilinger, D., and Howe, A. Experiences with selecting search engines using metasearchI. *ACM TOIS 15*, 3 (1997), 195–222.
6. Gauch, S., and Wang, H. Information fusion with ProFusion. In *Proceedings of the First World Conference of the Web Society* (San Francisco, CA, USA, 1996).
7. Gauch, S., Wang, H., and Gomez, M. ProFusion: Intelligent fusion from multiple distributed search engines. *Journal of Universal Computing 2*, 9 (1996).
8. Gravano, L., Chang, K., Garcia-Molina, H., and Paepcke, A. STARTS - Stanford protocol proposal for internet meta-searching. In *ACM SIGMOD* (1997).
9. Huibers, T. W. C. *An Axiomatic Theory for Information Retrieval*. PhD thesis, Utrecht University, The Netherlands, 1996.
10. Kantor, P., Ng, K., Hirsh, H., Basu, C., Loewenstern, D., and Kudenko, D. Data fusion of machine learning methods for the TREC-5 routing task (and other work). In *Proceedings of the fifth Text REtrieval Conference (TREC-5), NIST Special Publication* (1995).
11. Ruspini, E. H. The logical foundations of evidential reasoning. Tech. Rep. 408, SRI International, 1986.
12. Savoy, J., Calve, A. L., and Vrajitoru, D. Report on the TREC-5 experiment: Data fusion and collection fusion. In *Proceedings of the fifth Text REtrieval Conference (TREC-5), NIST Special Publication* (1995).
13. Selberg, E., and Etzioni, O. The MetaCrawler architecture for resource aggregation on the web. *IEEE Expert 12*, 1 (1997), 8–14.
14. Smeaton, A. Independence of contributing retrieval strategies in data fusion for effective information retrieval. In *Proceedings of the 20th BCS-IRSG Colloquium, Grenoble, France* (1998), Springer-Verlag Workshops in Computing. in press.
15. Smeaton, A., and Crimmins, F. Using a data fusion agent for searching the www. Poster presented at the WWW6 conference, 1997.
16. Tsikrika, T., and Lalmas, M. Merging techniques for performing data fusion on the web. In *Conference on Information and Knowledge Management (CIKM)* (2001), pp. 181–189.
17. Voorhees, E., Gupta, N., and Johnson-Laird, B. Learning collection fusion strategies. In *Proc. 18th ACM SIGIR Conference* (Seattle, USA, 1995), pp. 172–179.
18. Yager, R., and Rybalov, A. On the fusion of documents from multiple collection information retrieval systems. *JASIS 49*, 13 (1998), 1177–1184.