A Belief Model of Query Difficulty that uses Subjective Logic

Christina Lioma[1], Roi Blanco[2], Raquel Mochales Palau[1], and Marie-Francine Moens[1]

[1] Computer Science, Katholieke Universiteit Leuven, 3000, Belgium
[2] IRLab, Computer Science Department, A Coruña University, Spain
christina.lioma@cs.kuleuven.be, rblanco@udc.es,
raquel.mochales-palau@cs.kuleuven.be, sien.moens@cs.kuleuven.be

Abstract. The difficulty of a user query can affect the performance of Information Retrieval (IR) systems. This work presents a formal model for quantifying and reasoning about query difficulty as follows: Query difficulty is considered to be a subjective belief, which is formulated on the basis of various types of evidence. This allows us to define a belief model and a set of operators for combining evidence of query difficulty. The belief model uses *subjective logic*, a type of probabilistic logic for modeling uncertainties. An application of this model with semantic and pragmatic evidence about 150 TREC queries illustrates the potential flexibility of this framework in expressing and combining evidence. To our knowledge, this is the first application of subjective logic to IR.

1 Introduction

The task of an Information Retrieval (IR) system is to retrieve information from a large repository of data in response to a user need, or *query*. The difficulty of this task may be affected by various factors, relating to the system or algorithms used, to the properties of the data to be retrieved, or to the inherent difficulty of the user's information need. The effect of the last of these factors upon retrieval performance is often referred to as *query difficulty*, and is studied extensively in the field (discussed in Section 5). Our work addresses query difficulty by proposing a formal framework for modelling query difficulty. Our proposed formalisation consists of a belief model that considers query difficulty to be a subjective belief, which is formulated on the basis of different types of evidence. This belief model uses a type of logic called subjective logic [11] in order to combine this evidence and to make a final estimation about the expected difficulty of a query. Any type of evidence can be used with this model.

Subjective logic is a type of probabilistic logic that allows probability values to be expressed with degrees of uncertainty. Like any probabilistic logic, it combines the strengths of logic and probabilities: from the area of logic, it draws the capacity to express structured argument models, and from the area of probabilites it draws the power to express degrees of those arguments. This means that one can reason with argument models in the presence of uncertain or partially incomplete evidence. Since most of our knowledge or evidence about query difficulty in IR can never be complete, but rather tends to include degrees of uncertainty, subjective logic constitutes an appealing model for representing query difficulty, in the sense that the conclusions drawn reflect any ignorance and uncertainty of the input evidence.

Subjective logic is not the only formalism to model degrees of uncertainty. Several other mathematical models have been proposed to this end, the oldest being the Bayesian model of subjective probabilities (a survey of its foundations can be found in [8]). There also exist generalisations of the Bayesian model, (critically surveyed in [21]), the best-known of which is Dempster-Shafer's *belief theory* [7,19]. The point of departure for Dempster-Shafer from classical Bayesian theory is its abandoning of the additivity principle of classical probabilities, i.e. the requirement that in a given event space, the probabilities of mutually disjoint elements must add up to 1. In classical Bayesian theory, this requirement makes it necessary to estimate a probability value for every element of the event space, even though there might be no basis for it, for instance in the case of uncertainty. Instead, Dempster-Shafer's belief theory suggests assigning a so-called *belief mass* to the whole event space. This belief mass is defined on the basis of both evidence and uncertainty about the event, hence it constitutes a much more flexible way of representing beliefs than traditional probabilities. Subjective logic can be seen as an alternative to the Dempster-Shafer theory, its main difference from the former being in its definition and distribution of belief mass: subjective logic defines belief mass as a function of not only belief and uncertainty, but also of an apriori probability in the absence of any evidence; furthermore, subjective logic assigns this belief mass, not to the whole event space, but to the individual elements of the event space. It can be argued that this allows subjective logic to formulate more expressive beliefs than Dempster-Shafer theory [11].

One of the advantages of using a belief theory, be it with Dempster-Shafer theory or subjective logic, is that it allows to operate on the beliefs and fuse them. Fusing beliefs is a formal way of saying 'combining evidence'. In the context of IR, combining evidence is a process that aims to use different types of information that may enhance IR performance, but for which we have different degrees of uncertainty regarding the enhancement that they may bring [18]. In this work we use two different subjective logic operations to combine evidence about query difficulty: a fair *consensus* and a biased *recommendation* (also called *discounting*).

The contribution of this work lies in proposing a type of formal logic for IR, which has not been used before in the field, and in illustrating its application to the representation and combination of evidence about query difficulty. To our knowledge, whereas Dempster-Shafer theory has been used extensively in IR (see Section 5), this is the first application of subjective logic to IR.

In the rest of this paper, Section 2 introduces belief models for subjective logic and presents our proposed belief model of query difficulty. Section 3 introduces the subjective logic operators for combining evidence used in this work.

Section 4 illustrates the application of our proposed belief model of query difficulty with 150 TREC queries and different types of semantic and pragmatic evidence. Section 5 overviews related past work on logic models for IR and query difficulty. Section 6 summarises this work and suggests future research directions.

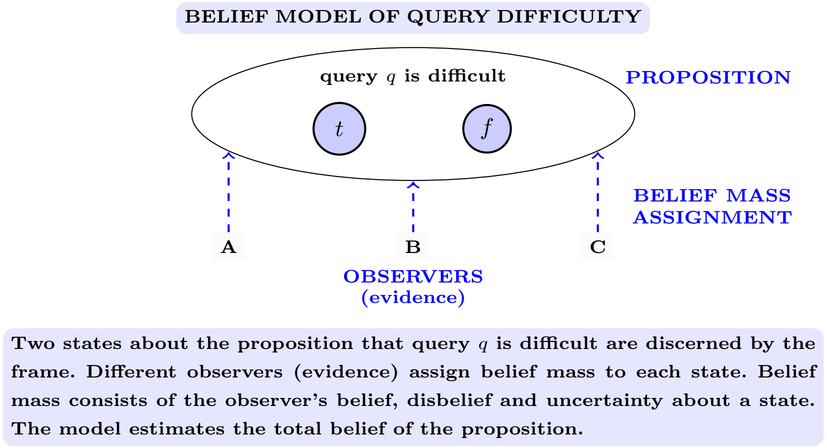
2 Belief Model of Query Difficulty

Belief models define a set of possible situations, for instance a set of possible states of a given system, called *frame of discernment*. This frame is defined over a proposition, i.e. a statement. At any one time, only one state of the frame of discernment can be true with respect to the proposition. A frame of discernment with two states ϕ and $\neg\phi$ is called *focused frame of discernment with focus on* ϕ. In this work, we use a focused frame of discernment.

Given any frame of discernment over a proposition, one can estimate the probability expectation that this proposition is true. This probability expectation is computed using evidence, which is said to come from 'observers'. An observer can assign to a state a *belief mass*, which represents his belief that this state is true with respect to the proposition. This belief can be represented in different ways by different uncertainty theories, for instance Dempster-Shafer or subjective logic as discussed in Section 1. An underlying similarity in these different representations is that this belief includes an explicit representation of the uncertainty of the observer about his belief.

Subjective logic considers the belief of an observer about the truth of a proposition as a subjective belief marked by degrees of uncertainty, and it calls it *opinion.* Let $\Phi = \{\phi, \neg\phi\}$ be a binary frame. An opinion about the truth of state ϕ is the ordered quadruple $\omega_\phi^A \equiv (b, d, u, a)$ where superscript A is the opinion's owner (i.e the **observer**), b is the belief mass supporting that the specified proposition is true (i.e. the **observer's belief**), d is the belief mass supporting that the specified proposition is false (i.e. the **observer's disbelief**), u is the amount of uncommitted belief mass (i.e. the **observer's uncertainty**), and a is the apriori probability in the absence of committed belief mass (divided uniformily among the states). These components satisfy: $b + d + u = 1$ and $b, d, u, a \in [0, 1]$. Clearly, a binomial opinion where $b + d = 1$ is equivalent to a traditional probability, and a binomial opinion where $b + d = 0$ expresses total uncertainty. The probability expectation of a binomial opinion is: $E = b + au$.

For the purpose of believing a binary proposition such as: "query q is difficult", we assume that the proposition will be either true or false. Hence, we define a focused frame of discernment as shown in Fig. 1 with the states: t (true) and f (false). The uncertainty probability of each state is represented by the belief mass assigned to each state by different observers, who are in fact our sources of evidence about query difficulty. The opinions of the three observers A, B, C shown in Fig. 1 are: $\omega_t^A \equiv (b_t^A, d_t^A, u_t^A, a_t^A)$, $\omega_f^A \equiv (b_f^A, d_f^A, u_f^A, a_f^A)$, $\omega_t^B \equiv (b_t^B, d_t^B, u_t^B, a_t^B)$, $\omega_f^B \equiv (b_f^B, d_f^B, u_f^B, a_f^B)$, $\omega_t^C \equiv (b_t^C, d_t^C, u_t^C, a_t^C)$, $\omega_f^C \equiv (b_f^C, d_f^C, u_f^C, a_f^C)$, where subscripts t, f denote the true and false state respectively. These opinions define our **opinion space**. The observers' opinions are drawn from real observations about the queries, which constitute our **evidence space**, discussed next.



Two states about the proposition that query q is difficult are discerned by the frame. Different observers (evidence) assign belief mass to each state. Belief mass consists of the observer's belief, disbelief and uncertainty about a state. The model estimates the total belief of the proposition.

Fig. 1. A belief model of query difficulty.

2.1 Evidence Space

For a focused frame of discernment, such as the one in Fig. 1, the proposition of the frame constitutes a binary event, where either the one or the other state is true: the query is either difficult or not. The type of evidence that we use to estimate the truth of this proposition can also be seen as binary, in the sense that it can be either positive (supporting that the query is difficult) or negative (supporting that the query is not difficult). Hence, both our opinion space and our evidence space consist of binary events. For such binary events, subjective logic defines a bijective mapping between the opinion and evidence space, as follows [11]. Let r denote positive evidence, and s denote negative evidence. Then, the correspondence between this evidence and the belief, disbelief, and uncertainty b, d, u is defined as:

$$b = \frac{r}{r+s+2} \qquad d = \frac{s}{r+s+2} \qquad u = \frac{2}{r+s+2} \tag{1}$$

Eq. 1 allows one to produce opinions based on statistical evidence. This mapping is derived in a mathematically elegant way, by considering the posteriori probability of the binary events defined in a focused frame of discernment, expressed using a beta probability function (the full derivation is presented in [11] and is outwith the focus of our work). The point to remember here is that in subjective logic any opinion has an equivalent mathematical and interpretative representation as a probability density function and vice versa.

3 Subjective Logic Operations for Combining Evidence

Subjective logic contains several operators for combining evidence (see [11] for more). In this work we use two combinations only: *consensus* and *recommendation* (or *discounting*). Using subjective logic terminology, we will refer to combining evidence as combining opinions, and treat these statements as equivalent.

3.1 Consensus between Independent Opinions

Let $\omega_x^A \equiv (b_x^A, d_x^A, u_x^A, a_x^A)$ and $\omega_x^B \equiv (b_x^B, d_x^B, u_x^B, a_x^B)$ be opinions respectively held by two independent observers A and B about the same proposition x. Then, $\omega_x^{A,B} \equiv (b_x^{A,B}, d_x^{A,B}, u_x^{A,B}, u_x^{A,B})$ is the opinion of an imaginary observer $[A, B]$ about x. $[A, B]$ represents the *Bayesian consensus* of opinions of both A and B, denoted $\omega_x^{A,B} = \omega_x^A \oplus \omega_x^b$, and defined by:

$$b_x^{A,B} = \frac{b_x^A u_x^B + b_x^B u_x^A}{\kappa}, \qquad d_x^{A,B} = \frac{d_x^A u_x^B + d_x^B u_x^A}{\kappa}, \qquad u_x^{A,B} = \frac{u_x^A u_x^B}{\kappa} \tag{2}$$

$$a_x^{A,B} = \frac{a_x^B u_x^A + a_x^A u_x^B - (a_x^A + a_x^B) u_x^A u_x^B}{u_x^A + u_x^B - 2 u_x^A u_x^B} \tag{3}$$

where $\kappa = u_x^A + u_x^B - u_x^A u_x^B$ such that $\kappa \neq 0$, and where $a_x^{A,B} = (a_x^A + a_x^B)/2$ when $u_x^A, u_x^B = 1$. The proof is included in [11].

This operation is both commutative and associative, meaning that the order in which opinions are combined does not impact the combination. The operation assumes that opinions are independent and that not all the combined opinions have zero uncertainty. Attempting to combine opinions all of which have zero uncertainty can be seen as meaningless, because these opinions would have complete belief or disbelief, and would hence be in complete conflict or agreement.

The effect of the consensus operator is to reduce uncertainty. The consensus operator has the same purpose as Dempster's rule [7], and the two tend to produce overall quite similar results. In [11], Section 5.3, Josang illustrates some cases where the consensus operator is 'better' than Dempster's rule, in the sense that the former produces less counterintuitive results than the latter.

3.2 Recommendation (or Discounting) between Opinions

Assume two observers A and B where A has an opinion about B, and B has an opinion about a proposition x. A recommendation of these two opinions consists of combining A's opinion about B with B's opinion about x[3] in order for A to get an opinion about x. Let $\omega_x^B \equiv (b_x^B, d_x^B, u_x^B, a_x^B)$ be B's opinion about x expressed in a recommendation to A, and let $\omega_B^A \equiv (b_B^A, d_B^A, u_B^A, a_B^A)$ be A's opinion about

[3] B's recommendation must be interpreted as what B recommends to A, and not necessarily as B's real opinion.

B's recommendation. Then, $\omega_x^{AB} = \omega_B^A \otimes \omega_x^B$ is A's opinion about x as a result of the recommendation from B, defined as:

$$b_x^{AB} = b_B^A b_x^B, \qquad d_x^{AB} = b_B^A d_x^B, \qquad u_x^{AB} = d_B^A + u_B^A + b_B^A u_x^B \qquad a_x^{AB} = a_x^B \quad (4)$$

This operation is associative but not commutative, meaning that the order in which opinions are combined impacts the combination. Eq. 4 can become equivalent to Shafer's discounting function [19] by setting $1-c = b_B^A$, where c denotes Shafer's discounting rate which is multiplied to the belief mass on each state in the frame except the belief mass of the powerset itself.

4 Illustrative Experiments

4.1 Evidence of Query Difficulty

We illustrate an application of our formalisation of query difficulty using two types of linguistic evidence, namely semantic and pragmatic evidence. From these types of linguistic evidence we obtain positive and negative evidence of query difficulty, which we map into belief, disbelief and uncertainty using Eq. 1.

The choice of evidence is illustrative. Our proposed model allows to represent and combine any type of evidence, simply by introducing more observers who contribute their beliefs to the frame of discernment. Any other evidence can be used.

Semantic Evidence. We use as semantic evidence two indicators that have been found to be correlated with query difficulty, namely (i) query scope, proposed by Plachouras and Ounis [18], and (ii) query polysemy [16]. Query scope is a probabilistic measure that estimates how specific or generic a query is by using the query term frequencies in the collection as well as their semantic content. Assuming that each query term corresponds to one or more concepts in WordNet (or any other similar lexical reference system), the semantic content of query terms is approximated from the hierarchical structure of their respective concepts. Specifically, we use the following formulae from [18] (keeping their original notation): given a term t_k and several concepts $\mathbf{C}_k$ associated to this term, the scope of t_k is defined as the maximum probability of any of its associated concepts: $scope_{t_k} = \max_{C \in \mathbf{C}_k} prob(C)$. This probability is estimated as follows: $prob(C) = \sum_{C=C_{k,j} \in \mathbf{C}_k} a_{k,j} \cdot \frac{tf_k}{T}$, where $C_{k,j}$ denotes the j^{th} concept (among all $\mathbf{C}_k$) associated to t_k, $a_{k,j}$ denotes the 'contribution' of t_k to concept $C_{k,j}$, tf_k is the frequency of t_k in the collection, and T is the sum of all the frequencies of all terms in the collection. The contribution $a_{k,j}$ is defined as: $a_{k,j} = \frac{(D_k+1)-d_{k,j}}{n_k(D_k+1)-\sum_{j=1}^{n_k} d_{k,j}}$, where $d_{k,j}$ denotes the length of the path from concept $C_{k,j}$ to the most generic concept in the WordNet hierarchy, D_k denotes the maximum path length of concepts $C_{k,j} \in \mathbf{C}_k$, and n_k denotes the number of concepts associated with t_k (in this work we select n_k among $C_{k,j}$ only). Plachouras and Ounis posit that as term scope approaches zero, the term is less represented in the collection, and hence more difficult to retrieve [18]. Based on this reasoning, we define a threshold θ_{sco}, so that any term scope $\leq \theta_{sco}$ constitutes positive evidence of query difficulty, and any term score $> \theta_{sco}$ constitutes negative evidence. For the illustrations shown here, we define θ_{sco} as the median term scope in all queries.

The second type of semantic evidence consists of the 'polysemy score' offered by WordNet to each term. This score reflects the number of concepts to which a term is associated, e.g. a score of 1 denotes a monosemous term. WordNet considers terms of polysemy score 1-4 as uncommon (in decreasing degrees from 1 to 4) and terms of polysemy score 5 or more as common (in increasing degrees from 5 upwards). Following the assumption that the more polysemous a term is, the harder the query [16], we define the following threshold: $\theta_{pol} \leq 4$ constitutes positive evidence of query difficulty, and $\theta_{pol} > 5$ constitutes negative evidence. Here, the value of the threshold θ_{pol} is taken directly from WordNet.

Pragmatic Evidence. Our pragmatic evidence aims to show whether a query constitutes a *literal* or *stipulative* statement of an information need. The meaning of a literal statement remains unchanged in all contexts, whereas the meaning of a stipulative statement is context- or register-dependent. We assume that a literal query should be easier to retrieve than a stipulative query because its meaning depends more on the literal semantics of its individual terms, and less on their contextual, metaphorical or other interpretations. To obtain this evidence, we use human judges, who read the queries and classify them as stipulative or not, based on their intuition. We use three human judges and consider their decision about the query being a literal or stipulative statement of an information need as negative or positive evidence of query difficulty respectively. The judges have a disagreement rate of 17.1%, and an inter-annotator agreement of $\kappa = 0.413$, measured using Cohen's κ, which indicates moderate agreement.

To recapitulate, we have three types of evidence (scope, polysemy, pragmatic judgement), which correspond to the observers of our model (Fig. 1). Next we illustrate how we formalise this evidence in our belief model of query difficulty, using TREC [24] queries.

4.2 Working Examples

Let us consider the following TREC queries: n^o 415 (`drugs, Golden Triangle`), n^o 479 (`where can I find information about kappa alpha psi?`), n^o 492 (`us savings bonds`), n^o 508 (`hair loss is a symptom of what diseases?`). In this section, we will estimate their difficulty and compare it to the retrieval performance they yield on their respective TREC collections (LAT & WT10G). Retrieval will be realised with the BM25 model at default settings, and measured by Mean Average Precision (MAP) against the prejudged relevance information provided for these datasets by TREC.

Table 1 presents the evidence given by our observers about the difficulty of each sample query, as well as the respective belief mass estimated by our model. The opinion of each observer can be represented as a tuple of the belief mass (b, d, u) and also of a prior probability of uncertainty a (a is divided uniformily among the two states of our frame of discernment, hence a=0.5 at all times).

Proposition: *the query is difficult*													
	query scope				**query polysemy**				**pragmatic judgement**				
query	belief	disbef.	uncert.	Exp.diff	belief	disbef.	uncert.	Exp.diff	belief	disbef.	uncert.	Exp.diff	MAP
	b	d	u	E	b	d	u	E	b	d	u	E	
415	0.13	0.63	0.25	0.25	0.40	0.20	0.40	0.50	0.60	0.00	0.40	0.80	0.250
479	0.50	0.33	0.17	0.58	0.29	0.43	0.29	0.44	0.60	0.00	0.40	0.80	0.240
492	0.75	0.13	0.13	0.88	0.20	0.40	0.40	0.30	0.60	0.00	0.40	0.80	0.307
508	0.79	0.11	0.11	0.84	0.33	0.33	0.33	0.50	0.00	0.60	0.40	0.20	0.230

Table 1. Sample queries with their respective query difficulty evidence (scope, polysemy, pragmatic judgment) and MAP. The belief mass components of each type of evidence are clearly shown (b, d, u), as well as their final expected probability of query difficulty (E).

For example, the opinion of the polysemy observer that query 492 is difficult is represented as $\omega_t^{pol} \equiv (0.4, 0.2, 0.4, 0.5)$. In this case, the observer's belief is equal to his uncertainty (=0.4), and his disbelief is low (=0.2), hence the evidence of this observer for this query is not very discriminative.

In Table 1 we also see that the different types of evidence do not always agree. For instance, using query scope evidence, the probability that query 415 is difficult is quite low (0.25), whereas using pragmatic evidence, the same probability for the same query is quite high (0.80). How difficult this query is can be seen in the last column of the table, which presents the MAP obtained by the system for this query. In this case, an MAP of 0.250 is relatively low, hence the pragmatic evidence seems more appropriate than the semantic scope evidence.

Combination of evidence								
	consensus sco,pol,pra				**pra discounts sco,pol**			
query	belief	disbef.	uncert.	Exp.diff	belief	disbef.	uncert.	Exp.diff
415	0.15	0.36	0.50	0.394	0.54	0.00	0.46	0.770
479	0.27	0.23	0.50	0.524	0.54	0.00	0.46	0.770
492	0.32	0.19	0.44	0.569	0.54	0.00	0.46	0.770
508	0.31	0.19	0.50	0.558	0.00	0.54	0.46	0.230

Table 2. Two different combinations of semantic scope (sco), polysemy (pol) and pragmatic (pra) evidence about the four sample queries: consensus combines all evidence fairly, whereas discounting favours pragmatic evidence at the expense of the other two.

A more accurate estimate of query difficulty could be obtained by combining those different types of evidence, as presented in Table 2. Table 2 illustrates the two combinations of evidence presented in Section 3. The column headed 'consensus' refers to the combination of our semantic scope, polysemy and pragmatic evidence (denoted 'sco,pol,pra' respectively) using Eq. 2. The column headed 'pra discounts sco,pol' refers to the combination by discounting (Eq. 4), where pragmatic evidence recommends its opinion to the consensus of scope and polysemy evidence. In this case, more weight is given to the pragmatic evidence than to the other two types of evidence. We see that combining evidence by consensus provides probability estimates that constitute a fair compromise of the individual expectations of each type of evidence. However, this is not always desirable, especially in cases such as the ones presented in Table 2, where the original estimates were in sharp disagreement between them. Combining strongly disagreeing evidence results in an expected probability that approaches 0.5, hence which can be considered relatively arbitrary. This implies that combining evidence by consensus may be better suited to generally agreeing evidence, than to sharply disagreeing evidence. On the contrary, combining evidence by discounting allows one to produce more biased estimates (in this context, bias is desirable). A prerequisite for such cases would be having some apriori knowledge regarding the reliability or suitability of each type of evidence, or about the agreement between the types of evidence to be combined. In a realistic situation, this type of knowledge is not difficult to obtain, since most systems that use query difficulty evidence for retrieval prediction ensure such knowledge using offline training and pre- or post-retrieval passes on prejudged relevant datasets (evidence relying on such processes in highlighted in Section 5). We see in Table 2 that combination by discounting produces estimates that are more discriminative than the consensus estimates, namely 0.77 and 0.23 as opposed to estimates closely approaching 0.5. The 0.77 estimates are in fact more accurate predictions of query difficulty, because the displayed queries are difficult queries (their MAP scores do not exceed 0.3, as shown in Table 1).

Finally, we can report that the observations reported illustratively above are also valid for the majority of the 401-550 TREC query set. Experiments with these 150 queries show that semantic scope is not discriminative evidence of query difficulty, that polysemy is better than scope but not at all times, and that pragmatic judgment constitutes the most reliable out of the three sources of evidence. More importantly, the combination of evidence for all queries is consistently better when we use discounting (with pragmatic evidence discounting the other two), than when we combine all three types of evidence on equal grounds with consensus. The respective correlation between the estimated query difficulty and MAP is in the range of Spearman's $\rho \approx 0.3$ for discounting (weak positive correlation), and $\rho \approx 0.1$ for consensus. These correlations are not strong, as is commonly reported for most types of query difficulty evidence, and in particular evidence stemming from the textual expression of the query [16]. These weak correlations are partly due to the reliability or quality of the evidence used, but also to the fact that the problem of query difficulty is largely influenced by several factors (as mentioned in Section 1), meaning that it is practically impossible for a single type of evidence to constitute a reliable and consistent predictor of query difficulty for all queries in all datasets [18].

5 Related Work

In order to avoid breaking the flow of the belief model presented in this work, we have left the treatment of related work at the end. This section discusses separately applications of formal logic to IR, and work on query difficulty. The applications of formal logic aim to give a plenary view of the different aspects and processes of an IR system that can be formalised with logic, hence constituting potential future applications of the subjective logic presented in this work. The overview of work on query difficulty aims to present different types of query difficulty evidence, which can be used within our proposed frame, using the same methodology and equations set out in Sections 2-3.

Formal Logic in IR. The expressive power of formal logic has long attracted applications of it to IR, starting with Van Rijsbergen's *logical uncertainty principle* [1]. Since then, modal logic has been used to integrate semantic-based and probabilistic-based approaches of deciding the relevance between a document and a query [17]. Extensions of the logical uncertainty principle have been proposed in order to integrate natural language processing and artificial intelligence techniques to IR [5]. Particular aspects of formal logic have also been used to address specific aspects or processes in IR, for instance belief revision has been used to model IR agents [14], to estimate the similarity between a document and a query [15], and more recently to model adaptive and context-sensitive IR [13]. The Dempster-Shafer theory presented in Section 1 has been used extensively: to build an IR framework where information structure, significance, uncertainty and partiality can be elegantly represented and processed [12], to integrate Web evidence into IR [23], to integrate into Web IR evidence of query difficulty in the form of semantic scope (one of the types of evidence we used in this work) [18], as well as to relate dependent indices [20]. There are further applications of formal logic to IR, reviews of which can be found in [3]. A more indepth treatment of formal representations for IR can be found in [2].

Query Difficulty. In this work we propose a formal representation of query difficulty, an area of much interest to IR. Difficult queries may be due to a number of causes. Linguistic features of the query text that may indicate query difficulty include morphological statistics (e.g. word length, number of morphemes per word), syntactical statistics (e.g. number of conjunctions, syntactic depth), or semantic statistics (e.g. polysemy value) [16]. Additional factors that may impact retrieval performance can be drawn from the retrieval resources. For instance, simple statistics such as the frequency of query terms in the collection [10], or the score of the top-ranked documents and the average inverse document frequency (idf) of query terms [22] have been correlated to query difficulty. Query difficulty has also been correlated with query length [25], based on the overlap between results of sub-queries based on single query terms and results of longer queries. A *clarity score* has been proposed [6] to measure the coherence of a list of retrieved documents by the KL-divergence between the query model and the collection model. A *robustness score* [26] has been proposed to quantify the robustness of the document ranking in the presence of uncertainty. Retrieval precision has been correlated to the distance between the retrieved document set and the collection [4] measured by the Jensen-Shannon divergence. In addition, different techniques have been proposed for predicting automatically query performance specifically in Web IR [27], either by making use of both single term and term proximity features to estimate the quality of top retrieved documents, or by viewing the retrieval system as a noisy channel, where the query is the input, the ranked list of documents is the corrupted output, and their proposed technique measures the degree of corruption. The main components of query difficulty have been defined as the textual expression of the query, the set of documents relevant to the query and the entire collection of documents, with experiments showing that query difficulty strongly depends on the distances between these components [4]. Finally, a recent overview of query difficulty with respect to performance prediction can be found in [9].

6 Conclusion

We proposed representing and formalising query difficulty for IR using subjective logic, a type of probabilistic logic for modelling uncertainties not used in IR before. Considering query difficulty as a subjective belief, formulated on the basis of various types of evidence, we defined a belief model that uses subjective logic, and a set of operators for combining evidence of query difficulty. We illustrated an application of this model with semantic and pragmatic evidence and TREC queries, which were combined in two different ways: by fair consensus and by biased discounting. Integrating further evidence or refining its combination can be realised easily with subjective logic, as illustrated in this work with working examples. Further research includes obtaining more varied sources of evidence for the task of query difficulty (any of the types of evidence highlighted in Section 5 can be used). Finally, the proposed belief model could be applied to other aspects of IR, apart from query difficulty, similarly to the varied and extensive use of Dempster-Shafer by the community (overviewed in Section 5).

Acknowledgements: Author 2 was partially supported by *Ministerio de Ciencia e Innovación*, FEDER and *Xunta de Galicia* under projects TIN2008-06566-C04-04 and 07SIN005206PR

References

1. C. J. van Rijsbergen. A non-classical logic for information retrieval. *Comput. J.*, 29(6):481–485, 1986.
2. C. J. van Rijsbergen. *The Geometry of Information Retrieval.* CUP, Cambridge, 2004.
3. C. J. van Rijsbergen, F. Crestani, and M. Lalmas. *Information Retrieval: Uncertainty and Logics.* Springer, 1998.
4. D. Carmel, E. Yom-Tov, A. Darlow, and D. Pelleg. What makes a query difficult? In *SIGIR*, pages 390–397, 2006.
5. Y. Chiaramella and J.-P. Chevallet. About retrieval models and logic. *Comput. J.*, 35(3):233–242, 1992.

6. S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *SIGIR*, pages 299–306, 2002.
7. A. P. Dempster. A generalization of Bayesian inference. *Journal of the Royal Statistical Society*, B(30):205–247, 1968.
8. P. C. Fishburn. The axioms of subjective probability. *Statistical Science*, 3(1):335–345, 1986.
9. C. Hauff, L. Azzopardi, and D. Hiemstra. The combination and evaluation of query performance prediction methods. In *ECIR*, pages 301–312, 2009.
10. B. He and I. Ounis. Inferring query performance using pre-retrieval predictors. In *SPIRE*, pages 43–54, 2004.
11. A. Josang. A logic for uncertain probabilities. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 9(3):279–311, 2001.
12. M. Lalmas. Information retrieval and Dempster-Shafer's theory of evidence. In *Applications of Uncertainty Formalisms*, pages 157–176, 1998.
13. R. Y. K. Lau, P. D. Bruza, and D. Song. Towards a belief-revision-based adaptive and context-sensitive information retrieval system. *ACM Trans. Inf. Syst.*, 26(2), 2008.
14. B. Logan, S. Reece, and K. Sparck Jones. Modelling information retrieval agents with belief revision. In *SIGIR*, pages 91–100, 1994.
15. D. E. Losada and A. Barreiro. A logical model for information retrieval based on propositional logic and belief revision. *Comput. J.*, 44(5):410–424, 2001.
16. J. Mothe and L. Tanguy. Linguistic features to predict query difficulty - a case study on previous TREC campaigns. In *SIGIR Workshop on Predicting Query Difficulty: Methods and Applications*, 2005.
17. J.-Y. Nie. Towards a probabilistic modal logic for semantic-based information retrieval. In *SIGIR*, pages 140–151, 1992.
18. V. Plachouras and I. Ounis. Dempster-Shafer theory for a query-biased combination of evidence on the web. *Inf. Retr.*, 8(2):197–218, 2005.
19. G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, 1976.
20. L. Shi, J.-Y. Nie, and G. Cao. Relating dependent indexes using Dempster-Shafer theory. In *CIKM*, pages 429–438, 2008.
21. P. Smets. *What is Dempster-Shafer's model?* Wiley, 1994.
22. S. Tomlinson. Robust, web, and terabyte retrieval with Hummingbird SearchServer at TREC 2004. In *TREC*, 2004.
23. T. Tsikrika and M. Lalmas. Combining evidence for web retrieval using the inference network model: an experimental study. *Inf. Process. Manage.*, 40(5):751–772, 2004.
24. E. M. Voorhees and D. K. Harman. *TREC: Experiment and Evaluation in Information Retrieval.* MIT Press, 2005.
25. E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In *SIGIR*, pages 512–519, 2005.
26. Y. Zhou and W. B. Croft. Ranking robustness: a novel framework to predict query performance. In *CIKM*, pages 567–574, 2006.
27. Y. Zhou and W. B. Croft. Query performance prediction in web search environments. In *SIGIR*, pages 543–550, 2007.