A Dempster-Shafer Model for Document Retrieval using Noun Phrases

Abstract

In this paper, we propose a document retrieval system based on natural language processing of documents and queries. We use single terms and term groups as indexing elements to represent documents and queries. The model is formally expressed within the Dempster-Shafer Theory of Evidence. We discuss in detail how we use this theory to represent a document collection, indexing elements, documents and queries. The retrieval function is derived directly from the underlying theory. We then present an implementation of the model. The experimental work carried out is reported last.

1 Introduction

Document retrieval (DR) systems focus on the problem of retrieving documents relevant to a user's information need represented as a query. In this work, we concentrate on text-based systems. Traditional DR models usually represent documents and queries as a set of keywords. Hence, they cannot handle the diversity of the human language. Their simplistic representational approaches cannot, for example, differentiate between a query on "high risk of intoxication" and a query on "risk of high intoxication".

The use of *Natural Language Processing* (NLP) in DR attempts to overcome such shortcomings [8, 9]. NLP techniques are based on the fact that the content of a document or a query is encoded in natural language. They aim to extract accurate linguistic structures that are then used to represent documents and queries, where linguistic structures can vary from noun-phrases to tagged sentences.

In this work, we use shallow NLP techniques to represent documents and queries. Shallow NLP techniques are not domain specific and can be applied to any document collection. Our model is formally expressed within the *Dempster-Shafer (D-S) Theory of Evidence* [7]. This is a theory of uncertainty that yields structural representations (a set of propositions and their associated beliefs) from the available evidence. The evidence here are the indexing elements that come from applying shallow NLP techniques on documents and queries.

The first section gives a brief introduction to the D-S theory. In the next section, we present the theoretical model. A description of the implementation of the model follows. The experimental work conducted is then reported followed by an evaluation of the results. Conclusions and future work are discussed in the last section.

2 Dempster-Shafer's theory of evidence

In this section, we describe the main definitions of the D-S theory. Only the main points of the theory as they apply to our model are given.

DEFINITION 2.1 Let Ω be a finite non-empty set of mutually exhaustive and exclusive events. The set Ω is called a *frame of discernment*.

DEFINITION 2.2 Let 2^{Ω} be the set of all subsets of the set Ω, including the empty set $\emptyset$ and Ω itself.

DEFINITION 2.3 Given a frame of discernment Ω, the function $m : 2^\Omega \mapsto [0, 1]$ is called a *basic probability assignment* (bpa) if:

$$m(\emptyset) = 0 \quad \text{and} \quad \sum_{A \in 2^\Omega} m(A) = 1$$

The bpa represents a source of evidence supporting various subsets A in 2^Ω with value, or "degree of support", $m(A)$. The subsets A of 2^Ω such that $m(A) > 0$ are called *focal elements*.

DEFINITION 2.4 Given a bpa $m : 2^\Omega \mapsto [0, 1]$, a function $\text{Bel} : 2^\Omega \mapsto [0, 1]$, called a *belief function* over Ω, is defined as:

$$\text{Bel}(A) = \sum_{B \subseteq A} m(B)$$

The value $\text{Bel}(A)$ quantifies the strength of the total belief committed to A. In contrast, $m(A)$ quantifies the exact belief committed to A.

A particular characteristic of the theory is that the belief of an event being x does not necessarily imply that the belief associated to the negation of the event is $1 - x$ (as it happens in probability theory). In the absence of any other evidence to support the negation of the event, the remaining belief is assigned to the frame of discernment (all the possible events), and represents the *uncommitted belief*.

2.1 A logical interpretation of the D-S theory of evidence

The D-S has a logical interpretation, which we use in this paper. The events defining the frame of discernment Ω can be considered as a set of *elementary propositions*. The *non-elementary propositions* are the propositions defined as the *disjunctions* of the elementary propositions. The set 2^Ω is the set of all propositions, whether elementary or not, including true $\top$ and false $\bot$.

EXAMPLE 2.1 Let e_0, e_1, e_2 be three elementary propositions of a frame of discernment Ω. Note that e_0, e_1, e_2 represent exhaustive and mutually exclusive events; hence $e_0 \vee e_1 \vee e_2 = \top$ and $e_0 \wedge e_1 \wedge e_2 = \bot$. In this case, $2^\Omega = \{\bot, e_0, e_1, e_2, e_0 \vee e_1, e_0 \vee e_2, e_1 \vee e_2, \top\}$.

Using the logical view of the D-S theory, the bpa is defined as $m : 2^\Omega \mapsto [0, 1]$ such that:

$$m(\bot) = 0 \quad \text{and} \quad \sum_{p \in 2^\Omega} m(p) = 1$$

Respectively, the belief function $\text{Bel} : 2^\Omega \mapsto [0, 1]$ becomes:

$$\text{Bel}(q) = \sum_{p \to q} m(p)$$

where $\to$ is the material implication of classical logic. Verbally the belief on q depends on the bpa of the propositions of the frame that imply q.

EXAMPLE 2.2 Following our previous example, let $m(e_0) = 0.3, m(e_0 \vee e_1) = 0.4$. The uncommitted belief is captured by assigning a bpa value to the truth proposition $\top$, $m(\top) = 0.3$. All other propositions of 2^Ω have null bpa value.

In this case, we obtain $Bel(e_0) = 0.3$ since $e_0 \to e_0$, $Bel(e_0 \vee e_1) = 0.3 + 0.4 = 0.7$ since $e_0 \to e_0 \vee e_1$ and $e_0 \vee e_1 \to e_0 \vee e_1$, and $Bel(e_2) = 0$ since no proposition with non-null bpa value implies the proposition e_2. It should be noted that $Bel(\top) = 1$ since all propositions imply the truth proposition.

3   Description of the Model

In this section we present our model. First we describe the indexing elements upon which the model is based (section 3.1). Based on this, we explain how the document collection is represented as a frame of discernment (section 3.2). We show how documents are represented within the frame (section 3.3). Finally, we describe the query representation (section 3.4) and the retrieval function (section 3.5).

3.1   Indexing elements

The representation of documents is based on the two following categories of indexing elements:

Single Terms: These are the standard single terms used in traditional DR systems.

Term Groups: These are groups of words derived from noun-phrases extracted from documents and queries. The NLP tool that extract the noun-phrases is described in Section 4.

We construct propositions based on word content only. Therefore, the ordering of the words constituting term groups is ignored. Our aim is not to obtain an exact representation of the extracted linguistic structures, but to use them to derive a more precise description of document and query content. However, the model proposed in this paper can be extended to include word ordering.

EXAMPLE 3.1 Suppose that the noun-phrase *"red wine"* has been extracted from a document. The corresponding term group is $\{red, wine\}$. This term group adds to the document representation a more precise description, that is that the document is about "*Red* AND *Wine*". This is equivalent to "*Wine* AND *Red*" since word ordering is ignored. Suppose that the analysis of the document does not yield the above term group, but two single terms *"red"*, and *"wine"*. This means that the document is about "*Red*", and "*Wine*", but not necessarily about "*Red* AND *Wine*". In practice, this means that the document is about "*Red* OR *Wine*"

3.2   Document collections as frames

First, we give some preliminary definitions of the indexed document collection.

DEFINITION 3.1 Let $\mathbb{C} = \{\mathbb{D}_1, \dots, \mathbb{D}_N\}$ be a document collection, where N is the number of documents. Let $\mathbb{S} = \{s_1, \dots, s_S\}$ be the set of single terms that appear in the document collection $\mathbb{C}$, where S is the number of single terms in the document collection. Also, let $\mathbb{G} = \{g_1, \dots, g_G\}$ be the set of term groups in the document collection, where $g_i \subseteq \mathbb{S}$.

DEFINITION 3.2 For a document $\mathbb{D}_i \in \mathbb{C}$, let $\mathbb{S}_i \subseteq \mathbb{S}$ be the set of single terms that appear in $\mathbb{D}_i$. Also let $\mathbb{G}_i \subseteq \mathbb{G}$ be the set of term groups that appear in $\mathbb{D}_i$.

EXAMPLE 3.2 Consider a document collection $\mathbb{C} = \{\mathbb{D}_1\}$ where the document $\mathbb{D}_1$ consists of the only sentence:

> "*At every stop of our long motorcycle trip, we were drinking red dry wine*".

The following set of single terms (underlined in the sentence) are obtained[1] for document $\mathbb{D}_1$, $\mathbb{S}_1 = \{$*stop, long, motorcycl, trip, drink, red, dry, wine*$\}$. Also the set of term groups for document $\mathbb{D}_1$ is $\mathbb{G}_1 = \{\{$*long, motorcycl, trip*$\}$, $\{$*red, dry wine*$\}\}$.

For a document collection $\mathbb{C}$, a frame of discernment Ω is constructed based on the set $\mathbb{S}$. The elements of the frame are defined as mutually exclusive propositions derived from a boolean combination generated from the set $\mathbb{S}$.

---

[1]After removal of stop words and stemming (see section 4).

DEFINITION 3.3 For the set of single terms $\mathbb{S} = \{s_1, \ldots, s_S\}$ of a document collection $\mathbb{C}$, all the 2^S boolean combinational elements are generated using the terms $s \in \mathbb{S}$, the negations ($\neg$) of these terms and the boolean conjunction ($\wedge$). These boolean elements represent the elementary propositions of the constructed frame Ω. It can be shown that the number of constructed elementary propositions is 2^S.

EXAMPLE 3.3 Let $\mathbb{S} = \{\text{Long, Wine, Red}\}$, and $s_1 = \text{Long}$, $s_2 = \text{Wine}$, and $s_3 = \text{Red}$. We obtain $2^3 = 8$ elementary propositions forming the frame of discernment Ω:

e_0	$\neg Red \wedge \neg Wine \wedge \neg Long$
e_1	$\neg Red \wedge \neg Wine \wedge \ \ Long$
e_2	$\neg Red \wedge \ \ Wine \wedge \neg Long$
e_3	$\neg Red \wedge \ \ Wine \wedge \ \ Long$
e_4	$Red \wedge \neg Wine \wedge \neg Long$
e_5	$Red \wedge \neg Wine \wedge \ \ Long$
e_6	$Red \wedge \ \ Wine \wedge \neg Long$
e_7	$Red \wedge \ \ Wine \wedge \ \ Long$

3.3 Document representation

Having the document collection modelled as a frame of discernment, the representation of documents is achieved through a set of focal elements (section 3.3.1) and the associated bpa (section 3.3.2) defined on the frame of discernment.

3.3.1 Focal and indexing elements

In the D-S theory, focal elements correspond to propositions for which there is positive evidence. Therefore, focal elements can be used as propositions modelling the indexing elements (single terms and term groups) of a document. For a document $\mathbb{D}_i \in \mathbb{C}$, they are defined upon the sets $\mathbb{S}_i$ and $\mathbb{G}_i$.

DEFINITION 3.4 Every single term $s_j \in \mathbb{S}_i$ of a document $\mathbb{D}_i \in \mathbb{C}$ defines a focal element, e.g. the proposition p_j. Furthermore, every term group $g_k \in \mathbb{G}_i$ also defines a focal element, the proposition $p_k = \bigwedge_l p_l$ where each p_l is the proposition associated to single term r_l for $r_l \in g_k$. Θ_i is defined as the set that includes all the propositions representing single terms and group terms of the document $\mathbb{D}_i$.

EXAMPLE 3.4 Let $\mathbb{D}_1$ be the document with $S_1 = \{\text{Long, Wine, Red}\}$ and $G_1 = \{\{\text{Red, Wine}\}\}$. The following propositions are the focal elements modelling the indexing elements of the document:

p_1	$Long$
p_2	$Wine$
p_3	Red
p_4	$Red \wedge Wine$

The propositions modelling indexing elements must be defined in terms of the elementary propositions defining the frame of discernment.

DEFINITION 3.5 A proposition is represented as the disjunction of elementary propositions as follows:

$$\forall p_j \in \Theta_i \quad p_j = \bigvee \{ e_k \in \Omega \mid e_k \rightarrow p_j \}$$

EXAMPLE 3.5 The proposition p_1 in example 3.4 is defined in terms of the elementary propositions defined in example 3.3 as $e_1 \lor e_3 \lor e_5 \lor e_7$. This comes from the following implications:

$$(e_1 = \neg Red \land \neg Wine \land Long) \rightarrow Long$$
$$(e_3 = \neg Red \land Wine \land Long) \rightarrow Long$$
$$(e_5 = Red \land \neg Wine \land Long) \rightarrow Long$$
$$(e_7 = Red \land Wine \land Long) \rightarrow Long$$

The proposition $p_4 = e_6 \lor e_7$ is defined as such because of the following implications:

$$(e_6 = Red \land Wine \land \neg Long) \rightarrow Red \land Wine$$
$$(e_7 = Red \land Wine \land Long) \rightarrow Red \land Wine$$

The frame of discernment along with the propositions modelling the indexing elements of the document $\mathbb{D}_1$ is shown schematically in Figure 1.
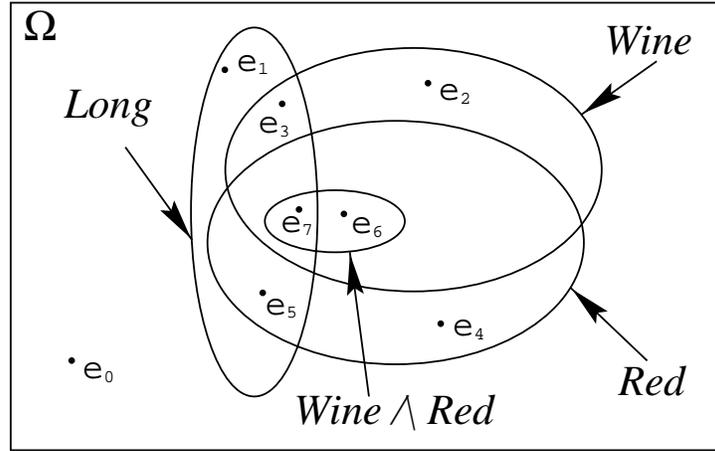

Figure 1: An example of a document in a frame of discernment

The truth proposition $\top$ can be viewed as the disjunction of all the elementary propositions. As explained in section 2.1, this proposition is used to capture the uncommitted belief for the document according to the D-S theory, and may then constitute a focal element.

The propositions used to model a document derive from observed evidence (the set $\mathbb{S}_i$ and $\mathbb{G}_i$ for document $\mathbb{D}_i$). Obviously, some propositions have stronger evidence than others. This is represented in the D-S via the use of a bpa.

3.3.2 Basic probability assignment

A bpa must be defined for every document $\mathbb{D}_i$ to capture the exact belief that the various propositions (focal elements) provide a good description of the document content. We compute the bpa values from term statistical characteristics in documents.

The bpa formula considered is:

$$m_i(p_j) = \begin{cases} \frac{\text{FREQ}_i(p_j)}{\text{TOTFREQ}_i} \cdot \text{IDF}(N, p_j) & p_j \in \Theta_i \\ \\ 0 & p_j \notin \Theta_i \text{ and } p_j \neq \top \end{cases} \tag{1}$$

where:

(i) $\mathrm{FREQ}_i(p_j)$ is the number of occurrences of the indexing element represented by the proposition p_j in the document $\mathbb{D}_i$.

(ii) $\mathrm{TOTFREQ}_i = \sum_{p_k \in \Theta_i} \mathrm{FREQ}_i(p_k)$ is the total number of occurrences of the indexing elements of the document $\mathbb{D}_i$.

(iii) $\mathrm{IDF}(N, p_j)$ is the inverted document frequency of the indexing element represented by the proposition p_j in a collection with N documents.

The first part of the formula ($p_j \in \Theta_i$) assigns a positive bpa value to propositions representing indexing elements of $\mathbb{S}_i$ and $\mathbb{G}_i$. The second part ($p_j \notin \Theta_i$) assigns 0 to all others propositions except for the truth proposition $\top$.

Verbally $\frac{\mathrm{FREQ}_i(p_j)}{\mathrm{TOTFREQ}_i}$ calculates term frequencies (i.e., occurrence number normalised by total occurrence). In [10] alternative variants of $\mathrm{FREQ}_i(p_j)$ were tested (e.g., different formulations were used to compute the frequency values of term groups) but failed to give better results than with the formulation above.

The remaining

$$1 - \sum_{p_k \in \Theta_i} m_i(p_k)$$

is treated as uncommitted belief and is assigned as the bpa value of the proposition $\top$. If non-null, $\top$ constitutes a focal element.

The various formulations of IDF in Formula (1) were motivated by the work of [3] on syntactic and statistical phrases:

$$\mathrm{IDF}(N, p_j) = \log_N \frac{N}{n(p_j)} \qquad \text{for } p_j \in \Theta_i \tag{2}$$

$$\mathrm{IDF}(N, p_j) = \begin{cases} \log_N \frac{N}{n(p_j)} & \text{for } p_j \text{ a single term} \\ \max_{p_k \in P_j} \left\{ \log_N \frac{N}{n(p_k)} \right\} & \text{for } p_j \text{ a term group} \end{cases} \tag{3}$$

$$\mathrm{IDF}(N, p_j) = \begin{cases} \log_N \frac{N}{n(p_j)} & \text{for } p_j \text{ a single term} \\ \mathrm{ave}_{p_k \in P_j} \left\{ \log_N \frac{N}{n(p_k)} \right\} & \text{for } p_j \text{ a term group} \end{cases} \tag{4}$$

$$\mathrm{IDF}(N, p_j) = \begin{cases} \log_N \frac{N}{n(p_j)} & \text{for } p_j \text{ a single term} \\ \min_{p_k \in P_j} \left\{ \log_N \frac{N}{n(p_k)} \right\} & \text{for } p_j \text{ a term group} \end{cases} \tag{5}$$

where:

(i) $n(p_k)$ is the number of documents that contain the indexing element represented by p_k.

(ii) P_i are the propositions p_ks that represent the single terms composing the term group p_i.

Formula (2) gives the standard IDF formula. In formulas (3), (4) and (5), the IDF value of single terms is as in Formula (2), whereas the IDF value for term groups is calculated as the maximum, the average and the minimum IDF values of the single terms that constitute the term groups, respectively.

Since the logarithms used in our formulas are on based N, the IDF values lie in the interval $[0, 1]$. As a result, the D-S restriction for the total bpa to be always equal to one ($\sum_{A \in 2^\Omega} m(A) = 1$) is satisfied (see [10]).

3.4   Query representation

Queries are represented as propositions defined in terms of the frame of discernment.

DEFINITION 3.6 Let $\mathbb{Q} = \{t_1, \ldots, t_q\}$ be the set of indexing elements used in a query. t_k can be a single term or a term group. To each query indexing element t_k we have an associated proposition q_k. The way the propositions are defined are the same as for documents (see section 3.3). For a single term that does not appear in the document collection (a term not defined in $\mathbb{S}$), the associated proposition is $\bot$.

DEFINITION 3.7 Let Q the set of propositions representing the query indexing elements of the set $\mathbb{Q}$. The query is represented by a proposition q defined as follows:

$$q = \bigvee_{q_k \in Q} q_k$$

The disjunction ($\vee$) is used since it is difficult to derive from a natural language query whether a user is seeking, for instance, documents about "*red wine*" or documents about "*red*" or "*wine*" unless the former is found as a term group in the query.

We consider two representations of queries:

Single term queries: Queries where only single term are used to express the query proposition q.

Term group queries: Queries which contain single terms and term groups. Single terms that appear only in term groups and not as stand-alone single terms in a query are represented in the query proposition only as a part of the term group proposition.

EXAMPLE 3.6 Consider the following query sentence:

"*Documents about red wine*".

Based on our previous example, if only single terms are considered, we obtain $\mathbb{Q}_1 = \{document, red, wine\}$ and the query proposition is $q = \bot \vee p_2 \vee p_3$. The term "*document*" is represented with the false proposition ($\bot$) because the term is not part of the set $\mathbb{S}$. If term groups are used, we have ($\mathbb{Q}_1 = \{document, \{red, wine\}\}$), and the query proposition becomes $q = \bot \vee p_4$.

3.5   Retrieval

To estimate the degree of relevance of a document to a query, we use the belief function of the D-S theory. To each document $\mathbb{D}_i$ with bpa m_i, we have an associated belief function Bel_i defined upon m_i. The degree of relevance of the document to a query represented by the proposition q is formulated as:

$$\mathrm{Bel}_i(q) = \sum_{p \to q} m_i(p)$$

This measure encapsulates the "relevance" of all the propositions used to describe the document content that imply the query formula q. If $\mathrm{Bel}_i(q) = 0$, the document is not relevant to the query. For a document collection, we use the belief values $\mathrm{Bel}_i(q)$ to rank the documents according to their estimated relevance to the query.

4 Implementation

We use a part-of-speech tagger and a noun-phrase extractor for the extraction of noun-phrases from the document collections and queries. The NLP software used in this implementation was designed and implemented at the Language Technology Group of the Human Communication Research Centre, at the University of Edinburgh [2]. The tagger achieves 96–98% accuracy if all the words in the text are found in the taggers lexicon, and 88–92% if unknown words appear in the text.

The DR system used for indexing and retrieval is based on a heavily modified version of the SIRE system [6]. The extracted single terms and term groups were filtered from stop words [11] and stemmed [4]. So a phrase like "*the long motorcycle trip*" yields the term group "*long motorcycl trip*". Noun phrases reduced to stand-alone terms because of stop word removal were not considered as noun-phrases, but as single terms.

5 Experiments and Evaluation

For the evaluation of the model, several experiments were performed. In this section we first describe the document collections used for the experiments followed by the model evaluation.

5.1 Document collections

The Cranfield-1400 collection was used first to perform a large number of initial experiments because it is small compared with the modern document collections (Wall Street Journal, Financial Times etc.). The final experiments were performed on the Financial Times (FT) collection.

5.1.1 The Cranfield-1400 collection

This collection contains 1400 aeronautical abstracts and 225 queries. Two empty documents can be found in the collection. Usually, these documents are ignored. In this work, they are considered valid and are represented with one focal element, the truth proposition $\top$ with bpa value of 1.

In Table 1 some statistics of the Cranfield-1400 document collection are shown. It can be seen that the Cranfield-1400 collection is rich in term groups.

Cranfield 1400	Number of documents (queries)	Single term average frequency/doc	Noun Phrase average frequency/doc
Documents	1400	1.67	1.16
Queries	225	1.04	0.88
	Average single term frequency	Average term groups frequency	Average document length
Documents	95.24	23.51	118.75
Queries	9.49	2.63	12.02

Table 1: Statistical characteristics of the Cranfield-1400 collection

5.1.2 The Financial Times (FT) collection

For the second set of experiments the Financial Times (FT) collection from TREC-5 was used. The FT collection contains 210158 articles summing up to roughly 600M bytes of text.

TREC queries (also referred to as TREC topics) are longer than the Cranfield-1400 queries. Three levels of topic details are defined. The title level can be viewed as the query entered manually to a system by the user. The description

level is an expansion of the title part in one sentence. In the narrative part the properties that the relevant documents must have are explained.

In Table 2 the statistics for the FT collection are shown. Compared with the Cranfield-1400 collection, FT has longer documents on average. Also the TREC topics are longer than the Cranfield-1400 queries but fewer (50 compared to 225).

FT	Number of documents (queries)	Single term average frequency/doc	Noun Phrase average frequency/doc
Documents	210158	1.54	1.02
Topics	50	1.37	0.73
	Average single term frequency	Average term groups frequency	Average document length
Documents	200.26	37.55	237.81
Queries	11.26	2.76	14.02

Table 2: Statistical characteristics of the FT collection and TREC topics

For the experiments performed in this work only the combination of title and the description part of the TREC topics was used. Some initial experiments performed (see [10]) showed narrative queries frequently contain query terms which easily retrieve non-relevant documents.

5.2 Model evaluation

We compare our model model with the vector space model (VSM) [5]. We used two variants of the VSM. The first variant uses the following weighting function[2]:

$$w_i(p_j) = \frac{\log_2 \mathrm{FREQ}_i(p_j + 1)}{\log_2 \mathrm{TOTFREQ}_i} \cdot \mathrm{IDF}(N, p_j)$$

It was reported in [1] that such a formula "can be safely used" for retrieval. This variant of the VSM is labeled VSM(1).

The second variant of the VSM uses the formula (1) with the IDF variant (2). We label this variant VSM(2). This model can be considered as a special case of the proposed model where only single terms are used as indexing elements. In this case, $\mathrm{Bel}_i(p_j) = m_i(p_j)$ holds for every proposition p_j of any document $\mathbb{D}_i$.

It must be noted that both VSM(1) and VSM(2) use single terms for the representation of documents and queries. The comparisons here are done using the the 'best' variant (in terms of retrieval effectiveness) of the Formula (1) for our model. This variant combines Formula (1) and the IDF factor (5).

5.2.1 Single term queries

The first comparison examines the results obtained with our model using the single term queries. The comparison can be seen in Table 3 for the Cranfield-1400 and the FT collections.

It can be seen that our model did not achieve higher precision results than the VSM. The difference in average precision between our model and VSM(2) is 2% for the Cranfield-1400 collection and 1% for the FT.

5.2.2 Term group queries

Here a comparison of the results obtained with the model using the term group queries is presented. The comparison can be seen in Table 4 for the Cranfield-1400 and the FT collections respectively.

---

[2] The original formula refers to terms instead of propositions. Here the formula is adapted to reflect the model's terminology.

Rcl	Precision			Rcl	Precision		
pts	VSM(1)	VSM(2)	Belief model	pts	VSM(1)	VSM(2)	Belief model
0.1	0.7922	0.7212	0.6776	0.1	0.5664	0.4118	0.3728
0.2	0.6909	0.5934	0.5698	0.2	0.3989	0.2353	0.2179
0.3	0.5807	0.5011	0.4683	0.3	0.2787	0.1610	0.1456
0.4	0.4921	0.4229	0.3965	0.4	0.1812	0.1077	0.0996
0.5	0.4257	0.3770	0.3476	0.5	0.1147	0.0717	0.0672
0.6	0.3376	0.3011	0.2787	0.6	0.0712	0.0468	0.0434
0.7	0.2484	0.2280	0.2174	0.7	0.0412	0.0289	0.0274
0.8	0.2020	0.1873	0.1764	0.8	0.0220	0.0156	0.0151
0.9	0.1504	0.1370	0.1249	0.9	0.0105	0.0073	0.0073
1.0	0.1302	0.1197	0.1104	1.0	0.0020	0.0020	0.0020
Average Precision				Average Precision			
	0.4050	0.3589	0.3368		0.1686	0.1088	0.0998

(a) Cranfield **(b)** FT

Table 3: Comparing the model (single terms queries) with the VSM

In these results there is a dramatic precision drop of the average precision of our model ($> 10\%$). This happens because some stand-alone query terms are now missing from the queries as they are components of term groups of the query. Subsequently, relevant documents where only the stand-alone terms exist are never retrieved. For example a query represented by the proposition $red \wedge wine$ does not retrieve a document about $wine$ because $wine \not\rightarrow red \wedge wine$.

6   Conclusions and Future Work

The model described in this paper, although theoretically sound, fails to achieve improvements versus the standard vector space model. However, the model leads to moderate results, thus suggesting that it is able to perform realistic retrieval. Substantial enhancements can be made to acquire better retrieval results. Some possible enhancements are discussed in the following sections.

6.1   Retrieval method

Although the theoretical framework supporting the model is sound, the use of the belief function as the retrieval function does not improve retrieval effectiveness compared with the single term VSM. This is because the belief function decreases precision when non-null bpa values are given to term groups.

The main reason is that single terms with low IDF value (i.e. frequent in the document collection) are also found in many noun-phrases, which poses a problem with single terms query. For example, in the FT collection the term "*account*" is a very frequent term, which also appears in many term groups (e.g., "*banking account*", "*saving account*", etc.). A query "*swiss account*" is represented as the proposition $swiss \vee account$. A document containing many noun-phrases with the term "*account*" is retrieved with high belief value even if the word "*swiss*" is not contained in the document.

Using group terms query solves this problem, but decreases recall. Only documents containing the noun-phrase "*swiss account*" (or longer noun-phrases including the terms "*swiss* and "*account*") are retrieved. The documents that have only the terms "*swiss*" or "*account*", but that do not contain the noun-phrase "*swiss account*" are not be retrieved.

One way to overcome this problem is to not use the belief function as the retrieval function. An alternative approach could be to use the Dempster's combination rule [7] for combining the document collection frame with a frame constructed from the query. This method requires a different representation of queries.

Rcl	Precision			Rcl	Precision		
pts	VSM(1)	VSM(2)	Belief model	pts	VSM(1)	VSM(2)	Belief model
0.1	0.7922	0.7212	0.3350	0.1	0.5664	0.4118	0.1745
0.2	0.6909	0.5934	0.2843	0.2	0.3989	0.2353	0.0836
0.3	0.5807	0.5011	0.1971	0.3	0.2778	0.1610	0.0465
0.4	0.4921	0.4229	0.1521	0.4	0.1812	0.1077	0.0283
0.5	0.4257	0.3770	0.1298	0.5	0.1147	0.0717	0.0189
0.6	0.3376	0.3011	0.0996	0.6	0.0712	0.0468	0.0130
0.7	0.2484	0.2280	0.0689	0.7	0.0412	0.0289	0.0085
0.8	0.2020	0.1873	0.0531	0.8	0.0220	0.0156	0.0053
0.9	0.1504	0.1370	0.0417	0.9	0.0105	0.0073	0.0033
1.0	0.1302	0.1197	0.0359	1.0	0.0020	0.0020	0.0019
Average Precision				Average Precision			
	0.4050	0.3589	0.1398		0.1686	0.1088	0.0384

(a) Cranfield-1400   **(b)** FT

Table 4: Comparing the model (term group queries) with the VSM

Also, uncommitted belief can be used to rank retrieved documents. Documents with low uncommitted belief should be ranked higher than those with high uncommitted belief. For instance, in the Cranfield-1400 collection the two empty documents will always be ranked last using this method. The appropriate way to incorporate uncommitted belief should be investigated.

6.1.1 Document length normalisation

The document length normalisation formula used in computing the bpa biases the retrieval of short documents. This happens also to VSM(2) where document length normalisation is also applied. Long documents relevant to a query are unable to compete with short documents. This phenomenon is amplified in the experiments with the FT collection. Since the number of documents in the FT collection is larger, it is more likely for a non-relevant short document to be highly ranked. Therefore, different methods for normalisation function that are compliant with the D-S theory must be exploited.

To sum up, the model presented in this paper uses shallow NLP techniques to obtain indexing elements to represent documents and queries. The model is based on a well-developed formal theory, the Dempster-Shafer Theory of Evidence. The various functions of the model are formally expressed within the theory. However, when implemented the model does not lead to better retrieval effectiveness than that of classical DR models. Nevertheless the model is at its infancy and there is plenty of space for improvements and enhancements, thus achieving effective retrieval results.

References

[1] D. K. Harman. Ranking algorithms. In W. B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures and Algorithms*, chapter 14, pages 363–392. Prentice-Hall, New Jersey, NJ, 1992.

[2] A. Mikheev and S. Finch. A workbench for finding structure in texts. In *Proceedings of the Applied Natural Language Processing (ANLP-97), Washington D.C.*, Apr. 1997.

[3] M. Mitra, C. Buckley, A. Singhal, and C. Cardie. An analysis of statistical and syntactic phrases. In *RIAO 97 Conference Proceedings*, volume 1, pages 200–214, June 1997.

[4] M. F. Porter. An algorithm for suffix stripping. *PROGRAM*, 14(3):130–137, 1980.

[5] G. Salton, A. Wong, and C. S. Yang. A vector space model for information retrieval. *Communications of the ACM*, 18(11):613–620, Nov. 1975.

[6] M. Sanderson. System for Information Retrieval Experiments (SIRE). Technical report, Department of Computing Science, Glasgow University, Glasgow G12 8QQ, Scotland, UK, Nov. 1996.

[7] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976.

[8] A. F. Smeaton. Progress in the application of natural language processing to information retrieval tasks. *The Computer Journal, Special Issue*, 36(3):268–278, 1992.

[9] A. F. Smeaton. Natural Language Processing and Information Retrieval. Tutorial notes presented at the Second European School in Information Retrieval (ESSIR 95), Glasgow, Scotland, Sept. 1995.

[10] M. Theophylactou. Document Retrieval using Natural Language Processing and the Dempster - Shafer Theory of Evidence. Master's thesis, University of Glasgow, Department of Computing Science, Sept. 1997.

[11] C. J. Van Rijsbergen. *Information Retrieval*. Butterworths, London, 2 edition, Jan. 1979.