Experimenting on Dempster-Shafer’s Theory of Evidence in Information Retrieval

Ian Ruthven and Mounia Lalmas
Department of Computing Science
University of Glasgow
Glasgow
G12 8QQ
email: igr, mounia@dcs.gla.ac.uk

April 27, 1998

Abstract

This report describes a set of experiments investigating the use of Dempster-Shafer’s Theory of Evidence in Information Retrieval. Our experiments use various indexing and retrieval methods to exploit Dempster-Shafer’s theory and we outline the reasons for the success or failure of the different approaches taken.

1 Introduction

The intention in writing this report is not to describe the theoretical advantages of using Dempster-Shafer’s (D-S) Theory of Evidence as a means of handling uncertainty in Information Retrieval (IR). The various arguments in favour of the Dempster-Shafer approach can be found in, for example [1, 2, 3, 4, 5, 6]. Rather, this report describes a set of experiments performed to test the various components of the D-S framework. Unlike other attempts to utilise this theory we have not relied on any external knowledge representation techniques, such as thesauri to manipulate knowledge or information about the retrieval process or the information content of documents.

The next section describes the features of Dempster-Shafer’s theory used in these experiments. The following section describes each of the experiments we carried out, the results and an analysis of the results. The final section is a discussion of our approach. The various test collections and standard weighting schemes used in these experiments are described in the Appendix. All these experiments used the Porter stemmer, [7], and the Van Rijsbergen stop word list, [8].

Table 4.7 summarises the experiments described in this report.
2 Dempster-Shafer’s Theory of Evidence

The Dempster-Shafer (D-S) framework is based on the view that propositions can be regarded as subsets of a given set of hypotheses\(^1\). For example, in IR, we can regard the set of hypotheses as the set of indexing terms of a collection. Each document, then, is a subset of \(U\). An example of a proposition is “the relevance of \(u\) is in \(A\)”. Thus, the propositions of interest are in a one-to-one correspondence with the subsets of \(U^2\), and the set of all propositions corresponds to the set of all subsets of \(U\), which is denoted \(2^U\). \(U\) is called a frame of discernment, and the propositions of interest are said to be discerned by the frame.

Beliefs can be assigned to propositions to express the uncertainty associated to them being discerned. The beliefs are usually computed based on a density function \(m : 2^U \rightarrow [0, 1]\) called a basic probability assignment (bpa):

\[
m(\emptyset) = 0 \quad \text{and} \quad \sum_{A \subseteq U} m(A) = 1.
\]

\(m(A)\) represents the belief exactly committed to the set \(A\). If \(m(A) > 0\), then \(A\) is called a focal element. The set of focal elements constitutes a core:

\[
C = \{ A \subseteq U \mid m(A) > 0 \}
\]

The core and its associated bpa define a body of evidence, from where a belief function \(Bel : 2^U \rightarrow [0, 1]\) is defined:

\[
Bel(A) = \sum_{B \subseteq A} m(B)
\]

\(Bel(A)\) is the total belief committed to \(A\), that is, the total positive effect the body of evidence has on the truth being in \(A\). The values \(Bel(A)\) can be interpreted as follows:

(i) \(Bel(A) = 0\) means that we have no knowledge about \(A\) or that \(A\) is false,

(ii) \(Bel(A) = 1\) means that \(A\) is true, and

(iii) \(0 < Bel(A) < 1\), means the evidence provides partial support for \(A\).

We can also talk of the plausibility of the evidence, employing a plausibility function \(Pl : 2^U \rightarrow [0, 1]\), defined as:

\[
Pl(A) = \sum_{B \cap A \neq \emptyset} m(B)
\]

The D-S theory has an operation, the Dempster’s rule of combination, for the pooling of evidence from a variety of sources. This rule aggregates two bodies of evidence defined within the same frame of discernment into one body of evidence. Let \(E_1 = (C_1, m_1)\) and \(E_2 = (C_2, m_2)\) be two bodies of evidence defined in \(U\). The new body of evidence is defined by \(E = (C, m)\), where the core \(C\) is:

\[^1\text{This description is taken from [9]}\]

\[^2\text{This also covers the freak case where two different documents are indexed by exactly the same set of index terms or keywords}\]
Any non-empty intersections of any two sets in the cores $C_1$ and $C_2$ is part of the new core $C$. These non-empty intersections reflect the case where the two bodies of evidence $E_1$ and $E_2$ agreed. The new bpa is:

$$m(A) = m_1 \oplus m_2(A) = \frac{\sum_{B \cap C = \emptyset} m_1(B)m_2(C)}{\sum_{B \cap C \neq \emptyset} m_1(B)m_2(C)}$$

In words, the Dempster’s combination rule computes a measure of agreement between two bodies of evidence concerning various propositions discerned from a common frame of discernment. The rule focuses only on those propositions that both bodies of evidence support. The new bpa takes into account the bpa associated to the propositions in $C_1$ and $C_2$ that yield the propositions of $C$. The denominator of the above equation is a normalisation factor that ensures that $m$ is a bpa (without it, there may be a belief associated to the empty set, which goes against the definition of a bpa).

### 3 Related work

There have been other attempts to use Dempster-Shafer’s Theory of Evidence in Information Retrieval, for example Schocken and Hummel, [5], used DS to combine taxonomies of keywords. In their work, human indexers were asked to assign sets of keywords, drawn from a fixed vocabulary, to a number of documents. They were also asked to assign a confidence level to each keyword set to represent their confidence in this set being a good description of the document. For example, for the document title "The Influence on Cezanne on the early work of Braque and Picasso", one indexer may select the keyword Cezanne with confidence level 0.6 and the set \{Braque, Picasso\} with confidence level 0.3. A different indexer may assign each keyword separately, \{Cezanne, 0.6\}, \{Braque, 0.15\}, \{Picasso, 0.15\}. The combination of these assignments, using Dempster’s combination rule, results in a new mass distribution over the sets that represents the combination of evidence from the two experts.

Schocken and Hummel advance this work by including taxonomic structures. For example, a document with the title "The Influence of Cezanne on the early work of Braque and Picasso" could have the description \{\{Braque, Picasso\}, \{Cezanne\} \}, and the document "The Influence of Cezanne on the early Cubists" could have the description \{\{Cubism\}, \{Cezanne\} \}. If we equate the set \{Picasso, Braque\} with the set \{Cubism\} then the both documents will have the same identifiers. If, on the other hand, we treat the set \{Braque, Picasso\} as being a subset of the set \{Cubism\} then any evidence for \{Picasso, Braque\} will support retrieval of the document "The Influence of Cezanne on the early Cubists" but not vice versa.

Although Schocken and Hummel’s paper is mainly directed at explaining and motivating the use of Dempster-Shafer’s Theory of Evidence in Information Retrieval, the reliance on a fixed vocabulary and pre-defined, domain-specific taxonomies make it difficult to assess the worth of DS in general, domain-independent IR.

An alternative use of DS is given by Jose and Harper, [10], who use DS as a means of data fusion in image retrieval. Each image is associated with a piece of text and with a
spatial representation. The spatial representation is given by manually assigning labels to marked areas of the image. A user may enter a textual query and/or a spatial query by sketching objects with given spatial relations. Each query component may be associated with a numerical confidence measure, giving his confidence in the component being a good indicator of his information need.

The images are ranked according to the similarity of their textual component to the user’s text query and separately ranked according to the similarity of the user’s spatial query to the image’s spatial description. This gives two sets of scores (source of evidence) over the same frame of discernment (the set of images) together with a measure of uncommitted belief for each source of evidence (1 - the confidence measure). DS is used to combine the rankings to give a ranking based on the two separate rankings.

4 Experiments

The experiments described in this report fall into two groups: the first group tests various aspects of DS theory and the second demonstrate an application of DS theory to structured document retrieval.

Experiments one to four are basic experiments to test the behaviour of the belief and plausibility functions in retrieving single, non-structured documents. We use the idf or tf*idf weighting schemes to calculate the bpa of each term in the collection (see Appendix). The use of the belief function here is identical to standard retrieval systems and in this experiment the belief function acts as our baseline retrieval performance.

4.1 Experiment one

Dempster-Shafer does not force all the evidence attached to a particular frame to be assigned to each hypothesis. Evidence can be retained, or a measure of uncertainty can be assigned to the whole event space. For example, Jose and Harper[10], allows users to assign a certainty measure to their query, describing how good the query is at representing their information need. In this experiment we treat each document as a frame we attach a value to each document In order to assess the effect of uncommitted belief of DS, we attached a belief value of 20

Results

The results from this experiment are shown in Figure 1. As can be seen the belief and plausibility recall-precision figures are identical. The explanation is fairly obvious: as we are adding a constant value to the score of each retrieved document, the rankings of retrieved documents for each query does not change. The belief and plausibility measures return identical

The results of experiment one mean that either each document must have a different level of uncertainty attached to them, or the identical terms appearing in different documents
should be weighted differently. The latter option suggests that the context in which terms appears should affect their weight.

Experiments two and three investigate altering the uncertainty attached to each document and experiments four and five investigate altering the weighting of terms in the document.

4.2 Experiment two

Experiment one attached a fixed uncertainty value to each document. This gave identical belief and plausibility rankings. In this experiment we varied the uncertainty value attached to each document. One way of interpreting the uncommitted belief value is to regard it as the degree to which the evidence for the terms in the document can improve. We calculate the uncertainty attached to each document by the equation.

\[
uncertainty_c = \sum_{t \in T} m_c(t) - \sum_{t \in c} m_c(t)
\]

In other words the degree to which the terms in the document can improve is equal to the sum of the terms not in the document.

Results

The results from this experiment are shown in Figure 2. The belief benchmark is identical to the one from experiment one. As can be seen the plausibility measure has plummeted in effectiveness. The main reason for this is that our method of calculating the uncommitted belief for each document unnaturally biases retrieval of short documents. The value of uncertainty is so large compared to the retrieval score of any document that we are effectively ranking documents by their uncertainty which is directly related to their length.
4.3 Experiment four

Results

4.4 Experiment five

Results

4.5 Experiment six

Results

Figure 2:

Figure 3:
4.6 Experiment seven

Results

4.7 Experiment eight

Results

<table>
<thead>
<tr>
<th>Exp</th>
<th>Bel</th>
<th>Pl</th>
<th>Norm</th>
<th>WtS</th>
<th>Uner</th>
<th>Struc</th>
<th>Coll</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>None</td>
<td>idf</td>
<td>Yes</td>
<td>No</td>
<td>Cacm/Npl</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>None</td>
<td>idf/tf*idf</td>
<td>Yes</td>
<td>No</td>
<td>Cacm</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>Mounia</td>
<td>idf</td>
<td>Yes</td>
<td>No</td>
<td>Cacm</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Yes</td>
<td>Ian</td>
<td>idf</td>
<td>Yes</td>
<td>No</td>
<td>Cacm</td>
</tr>
<tr>
<td>5</td>
<td>Yes</td>
<td>Yes</td>
<td>Doc mod</td>
<td>idf</td>
<td>Yes</td>
<td>No</td>
<td>Cacm</td>
</tr>
<tr>
<td>6</td>
<td>Yes</td>
<td>No</td>
<td>None</td>
<td>idf</td>
<td>No</td>
<td>Yes</td>
<td>Cacm/Cran</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>No</td>
<td>None</td>
<td>tf*idf</td>
<td>No</td>
<td>Yes</td>
<td>Cacm</td>
</tr>
<tr>
<td>8</td>
<td>Yes</td>
<td>No</td>
<td>Doc mod</td>
<td>idf</td>
<td>No</td>
<td>Yes</td>
<td>Cacm</td>
</tr>
</tbody>
</table>
Key: Bel = Measured the belief function,  
P1 = Measured the plausibility function,  
Norm = type of normalisation used if any,  
WtS = weighting scheme used,  
Uncer = used any notion of uncommitted belief,  
Struc = dealt with structured documents,  
Coll = collections used  
Mounia = sum terms in document such that doc terms = 1 - uncertainty,  
Ian = uncertainty attached to each doc = all terms - doc terms,  
Doc mod = adding terms to documents to reach an average document weight

5 Appendix

<table>
<thead>
<tr>
<th></th>
<th>Cacm</th>
<th>Cisi</th>
<th>Cranfield</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td>3204</td>
<td>1460</td>
<td>1400</td>
</tr>
<tr>
<td>Queries</td>
<td>64</td>
<td>111</td>
<td>225</td>
</tr>
<tr>
<td>Queries with relevance assessments</td>
<td>52</td>
<td>76</td>
<td>225</td>
</tr>
<tr>
<td>Unique terms in the collection</td>
<td>7816</td>
<td>7108</td>
<td>6336</td>
</tr>
<tr>
<td>Average document length</td>
<td>32</td>
<td>52</td>
<td>66</td>
</tr>
<tr>
<td>Average query length (all terms)</td>
<td>12</td>
<td>29</td>
<td>9</td>
</tr>
<tr>
<td>Average query length (terms in collection)</td>
<td>10</td>
<td>27</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 1: Details of the three standard test collections
References


