Information Retrieval using Dempster-Shafer Theory

Phuke V. A
Department of Computer Engineering, K. J.
Somaiya College of Engineering
Vidyanagar, Vidyavihar(E), Mumbai - 400 077

H. N. Bharathi
Department of Computer Engineering, K. J.
Somaiya College of Engineering
Vidyanagar, Vidyavihar(E), Mumbai - 400 077

ABSTRACT

Information retrieval model focuses on the problem of retrieving documents relevant to a user's information need represented as a query. One of the major difficulty of information retrieval is to find the relevance of documents with respect to the user query or information need. The choice of similarity measure is decisive for improving search effectiveness of a IR model. Different similarity measures have been proposed to find most relevant documents with the given query. Vector space model is a popular model and is widely used for information retrieval. The judgment of the relevance between a query and a document is evaluated using cosine similarity between them. However, vector space model does not give reasonable results in terms of precision and recall value.

Information retrieval model using Dempster-Shafer theory also known as evidence theory is used in this paper. In this model, each query-document pair is taken as a piece of evidence for the relevance between a document and a query. The evidence is combined using Dempster's rule of combination and the belief committed to the relevance is obtained which then ranked accordingly. To validate the feasibility of this approach, evidences for sample document collection i.e. TREC-9 filtering track i.e. OSHUMED dataset are considered and the results are compared with traditional VSM model in terms of precision and recall measures. It is found that Dempster Shafer Model's performance is better than VSM for information retrieval.

General Terms

Dempster Shafer Theory, Basic Probability Assignment, OSHUMED Dataset, Similarity Measure, Information Retrieval Model,

Keywords

Information Retrieval, Dempster Shafer Theory, Evidence Combination, Vector Space Model, Lucene.

1. INTRODUCTION

Information retrieval model is widely used as a system to effectively retrieve documents matching to a given query from the collection of documents. Many information retrieval models have been suggested and the vector space model (VSM) is a well-known among them. In VSM, document indexing is done initially where content bearing terms are extracted from the document. Subsequently weighting of the indexed terms is done to enhance retrieval of relevant documents. Finally ranking is done according to cosine similarity measure. Though VSM has been widely used for information retrieval, it results in decrease in recall and precision [1].

From a multi-source information collaboration point of view, a variety of research concerning the usage of the Dempster-Shafer theory of evidence in ranking has been conducted [2-4].The Dempster-Shafer theory of evidence is utilized to retrieve documents using noun phrases [2]. Single terms and term sets are taken as indexing terms, and belief functions are used to rank documents [3]. Content and structure information is combined using Dempster's rule of combination [4]. Even different sentences in a document are deemed to be multi-source information and are combined to help ranking [5].

Model used in this work is based on Dempster Shafer Theory model of Information Retrieval proposed by Zhang et al [6]. In this model, the particle of the multi-source evidence is reduced to each query-document pair, rather than the term sets, sentences, contents, or structures. Each $\{q_i ,t_j \}$ is taken as a piece of evidence that supports whether a document is relevant or irrelevant. This relevancy or irrelevancy of document to a query depend on $q_i = t_j$ or $q_i \neq t_j$. Whole collection of documents are taken into consideration. Multi-source evidence is combined by using Dempster's rule of combination to derive the belief committed to the relevance, which is then used to rank the documents.

Evidence Theory is a generalization of the Bayesian Model for Uncertainty that allows the notion of partial belief. The basic probability assignment function (also known as the mass function) assigns belief to sets of propositions rather than just singletons.

Dempster-Shafer theory can be applied to information retrieval model because of several reasons: First is that, D-S theory incorporates the uncertain nature of information retrieval. Second reason is that this evidence theory allows a degree of Belief to be associated with Ignorance. Third reason is that multiple evidence coming from different sources can be using Dempster's rule of combination.

The aim of this paper to implement retrieval model based on this Evidence Theory and then by going beyond i.e. simple information retrieval for TREC database and evaluate the results of retrieval in terms of precision and recall. In addition to that, evaluated results will be compared with the results of traditional VSM model for same document retrieval. In the end, this work gives an evaluation of the Dempster-Shafer's evidence theory and its potential applicability in the area of information retrieval.

2. DEMPSTER SHAFER THEORY FOR INFORMATION RETRIEVAL

Dempster theory of evidence is extension of basic probability theory. This theory can model both aleatory and epistemic uncertainties. Aleatory uncertainty arises because of randomness in the basic process. This type uncertainty can not be reduced as inherent to the process. The epistemic uncertainty arises because of lack of knowledge. This type uncertainty can be reduced by conducting more experiments. Dempster Shafer Theory is applied where both type of uncertainties exist together. As this theory can model uncertainties it can be applied to information retrieval.

2.1 Information retrieval

Information retrieval involves the goal of representing and storing information objects in a way that allows users to easily access the information that satisfies their respective information needs. Detailed view of the architecture of an IR system is shown in figure 1. Text preprocessing is done to remove multiple forms of words and convert them into generalized form using techniques such as stemming, stop word removal etc. This is followed by indexing. Indexing module constructs an inverted index from words to document pointers. The searching module retrieves documents that matches with users query, using the inverted index. According to the different similarity measures ranking module gives scores to all retrieved documents. The user interface accepts input in the form of users query and gives output as ranked documents. It includes the visualization of results. The query operations are used to refine the query so as to improve retrieval.

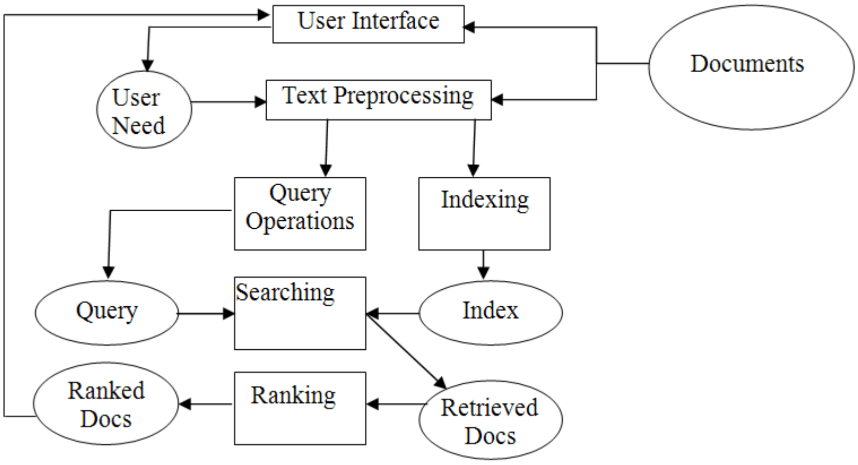

Fig. 1 Information retrieval model

Performance of the retrieval model is measured in terms recall and precision values. Precision is defined as the ratio of the number of relevant retrieved documents to the total number of retrieved documents [7].Recall is defined as ratio of number of relevant retrieved documents to the total number of relevant documents.

2.2 Fundamentals of Dempster-Shafer Theory (DST)

Dempster-Shafer's (D-S) Theory of Evidence is a theory of uncertainty that was first developed by Dempster and later extended by Shafer. This theory may be considered as a generalization of the probability theory. There are two differences with probability theory: First is the possibility of explicit representation of ignorance and second is the combination rule of evidences from multiple sources. Combination rule makes the D-S theory attractive for information retrieval process. Dempster's combination rule allows the expression of aggregation necessary in a model for structured document representation and retrieval. Other theories of uncertainty such as probability theory or possibility theory, the combination of evidence is not possible explicitly [8].

Basic building block of Dempster-Shafer theory is mass function m. This mass function is also called as basic belief assignment (bba) or basic probability assignment (bpa). Generally the term “basic probability assignment” does not refer to probability in the classical sense. The value of the bpa for a given set A (represented as m(A)), expresses the proportion of all relevant and available evidences that supports the claim that a particular element of Ω (the universal set) belongs to the set A but to no particular subset of A. The value of m(A) pertains only to the set A and makes no additional claims about any subsets of A. Any further evidence on the subsets of A would be represented by another bpa, m(B) would the bpa for the subset B.

2.3 Dempster’s Rule of Combination

If, m_1 and m_2, are two mass functions taken from two different sources. Each source is considered as evidence. These evidences can be combined by Dempster’s rule of combination resulting into a new mass function [8]

$$m(C) = \frac{\sum_{\substack{i,j \\ A_i \cap B_j = C}} m_1(A_i) m_2(B_j)}{1-K} \qquad (1)$$

where $K = \sum_{\substack{i,j \\ A_i \cap B_j = \phi}} m_1(A_i) m_2(B_j)$

where K represents basic probability associated with conflict.

Dempster’s rule of combination is very powerful tool in calculating the belief and plausibility values when we have multiple opinions or evidences from multiple sources.

3. APPLICATION OF DST TO IR

Our approach combines the DST model and pure Boolean models. The Boolean model needs a query in conjunctive normal form. Then it simply returns a set of matching documents. There is no relevance scoring; the result is simply match, or no match. In a DST model, each query-document pair is taken as a piece of evidence for the relevance between a document and a query. The evidence is combined using Dempster’s rule of combination, and the belief committed to the relevance is obtained. Retrieved documents are then ranked according to the belief committed to the relevance. To implement and validate model TREC-9 query is used as a input to the given TREC corpus. Retrieved documents by Boolean are ranked in descending score computed by using

DST-BPA. Finally the results (ranked list) are compared with predefined relevant judgments given in TREC dataset collection.

3.1 Evidence formulation

The aim is to retrieve the documents and rank them according to their relevance to the query. Let the frame of discernment be Θ = {R, $\bar{R}$} [6]. R means that the document is relevant to the query, while $\bar{R}$ means that the document is irrelevant to the query. The power set of Θ is P(Θ) = {∅,{R},{$\bar{R}$},Θ}. The space Θ denotes ignorance. Here we do have information of relevancy or irrelevancy. This ignorance arises because of lack of knowledge. Let the document have N terms and be expressed as D= ($t_1,t_2,..,t_N$) .Similarly, the query has M terms and is expressed as Q = ($q_1,q_2,..,q_M$).Each query-document pair is taken as a piece of evidence; in other words, {q_i,t_j} are taken as a piece of Evidence. They are indexed by i and j . From each piece of evidence, the extent to which the document is relevant to the query can be determined. A basic probability value may be assigned to each of them. Then, mass functions can be obtained. If $q_i = t_j$ then it indicates that the document is possibly relevant to the query. Similarly if $q_i \neq t_j$, then document cannot prove the relevance between a document and a query. It can also unjustified the relevance. Different assignments for mass function are proposed by Zang et al [6]. Additional mass function is also proposed in this paper.

3.2 Propositions for Basic Probability Assignments

The basic principle of choosing the assignment is that the appearance of a query-document pair {q_i,t_j } should support the fact that the document and the query are relevant. The occurrence of {$q_i = t_j$} ($q_i \neq t_j$)should support the irrelevance between the two, or at least should not give more support to the relevance between them. If the collection information is taken into account, the degree to which it should be reflected and quantified must also be determined.

Zang et al [6] has proposed many BPAs for information retrieval process.

Assignment 6 which uses both intra and inter document information along with Dempster has better performance than VSM and is considered for study. In this case mass value assigned is as follows:

$$m_{ij}(R) = \begin{cases} \dfrac{1}{M \times N}, & if\ q_i = t_j \\ 0, & if\ q_i \neq t_j \end{cases}$$

$$m_{ij}(\overline{R}) = \begin{cases} 0, & if\ q_i = t_j \\ \dfrac{1}{M \times N}, & if\ q_i \neq t_j \end{cases}$$

$$m_{ij}(\Theta) = 1 - \frac{1}{M \times N} \qquad (2)$$

Evidence are evaluated using above equations and are combined for each query term using Dempster's rule of combination.

Query terms collection is information is taken as another source of evidence . The assignment for the collection evidence of term i is:

$$m_{iC}(\mathrm{R}) = \frac{1}{N}\log_{\mathrm{C}}\left(\frac{\mathrm{C}}{\mathrm{n_i}}\right) \qquad (3)$$

Finally evidences for different query terms are combined by using Dempster's rule of combination.

In this paper additionally another mass function is suggested for query term collection. This assignment for query term collection is

$$m_{iC}(\mathrm{R}) = \left(\frac{1}{N}\log_{\mathrm{C}}\left(\frac{\mathrm{C}}{\mathrm{n_i}}\right)\right)^2 \qquad (4)$$

This is referred as basic belief assignment 7. Its performance with VSM and BPA 6 are compared in this work.

4. IMPLEMENTATION & RESULTS

Lucene open source code along with Netbeans IDE is used in this project for implementing VSM and DST model. MATLAB is used for fragmentation of large data collection and precision/recall measures of VSM and DST models.

4.1 Working with Lucene

Ranking of the document depends on score of the document that document gets. Lucene provides a similarity class which changing Similarity is an easy way to influence scoring, this is done at index-time with Similarity class and at query-time with IndexSearcher.setSimilarity (Similarity)[9]. Scoring can be influenced by configuring a different built-in Similarity implementation, or by tweaking its parameters, subclassing it to override behavior. Some implementations also offer a modular API which we can extend by plugging in a different component (e.g. term frequency normalizer).

Lucene provides Similarity measure which can be readily used. However in some applications it may be necessary to customize Lucene's inbuilt similarity implementation. To change Similarity, one must change indexing as well as searching. These changes must be done before either of these actions take place.

To make this change, implement own Similarity and then register the new class by calling IndexWriterConfig.setSimilarity(Similarity) before indexing and IndexSearcher.setSimilarity (Similarity) before searching. In Lucene, we have implemented new ranking method by extending Similarity Base, which provides basic implementations.

4.2 Evaluation of DST and VSM using TREC-9 (filtering track) Dataset

Text corpus used to evaluate these algorithms of VSM and DST is TREC-9 filtering track. The filtering track is to retrieve those document streams that match with user's interest as represented by query. The main focus of this track was on adaptive filtering track, for each user profile, the system begins with two identified relevant documents and a natural language description of the information need, which is the title and the description field of the corresponding topics provided by NIST. [10]

4.3 OSHUMED Data Collection

Document collection used for TREC-9 filtering track was OSHUMED dataset. The OHSUMED test collection is a set of 348,566 references from MEDLINE, the on-line medical information database, consisting of abstracts from 270 medical journals over a five-year period (1987-1991). The available fields are title, abstract, MeSH indexing terms, author, source, and publication type. The OHSUMED document collection was obtained by William Hersh and colleagues for the experiments. 63 OHSUMED queries were used to simulate user profiles. The relevance judgments were made by medical librarians and physicians based on the results of interactive searches.

There were three different sets of filtering topics for the TREC-9 Filtering track: one is the subset of 63 of the original query set developed by Hersh et al. for their IR experiments (OHSUMED),

Second is the set of 4904 MeSH terms and their definitions (MSH), and third is the subset of 500 of the MeSH terms (MSH-SMP).

OSHUMED data collection consists of Titles containing abstracts from medical journals, MESH queries (topics) asked to provide a statement of information about patient as well as their information need containing the full set of 4904 topics and set of relevant judgments that gives relevance and irrelevance of document with respect to query. [10]

4.4 Results

In this section, the experimental results of our evidential retrieval model are given. In following section, detailed description of results regarding work done is given.

VSM is implemented & validated its result with TREC9-OSHUMED filtering track dataset. For example MESH query 299 i.e. "Antigens Surface" taken as a query statement and calculated scores of each document related to given query MESH 299. By using cosine similarity measure, score calculated to rank the documents. BPA-6 and BPA-7 similarity measure explained 3.2 are also implemented in Lucene.

Precision and Recall of VSM, BPA-6 and proposed DST model BPA-7 are given in table 1 are also given in fig. 2. DST-BPA gives better performance compared to the simple VSM model.

Table 1. Precision and Recall Values

Recall	Precision		
	VSM	BPA-6	BPA-7
0.01	0.273	0.333	0.353
0.10	0.162	0.192	0.195
0.20	0.158	0.160	0.163
0.30	0.144	0.163	0.165
0.40	0.126	0.172	0.174
0.50	0.112	0.161	0.163
0.60	0.108	0.157	0.159
0.72	0.092	0.148	0.151
0.80	0.075	0.149	0.152
0.90	0.056	0.150	0.152
1.00	0.029	0.143	0.145

BPA 6 and BPA 7 by adding another similarity measure to LUCENE framework. Results of precision and recall for the experimental work done plotted in fig 2.

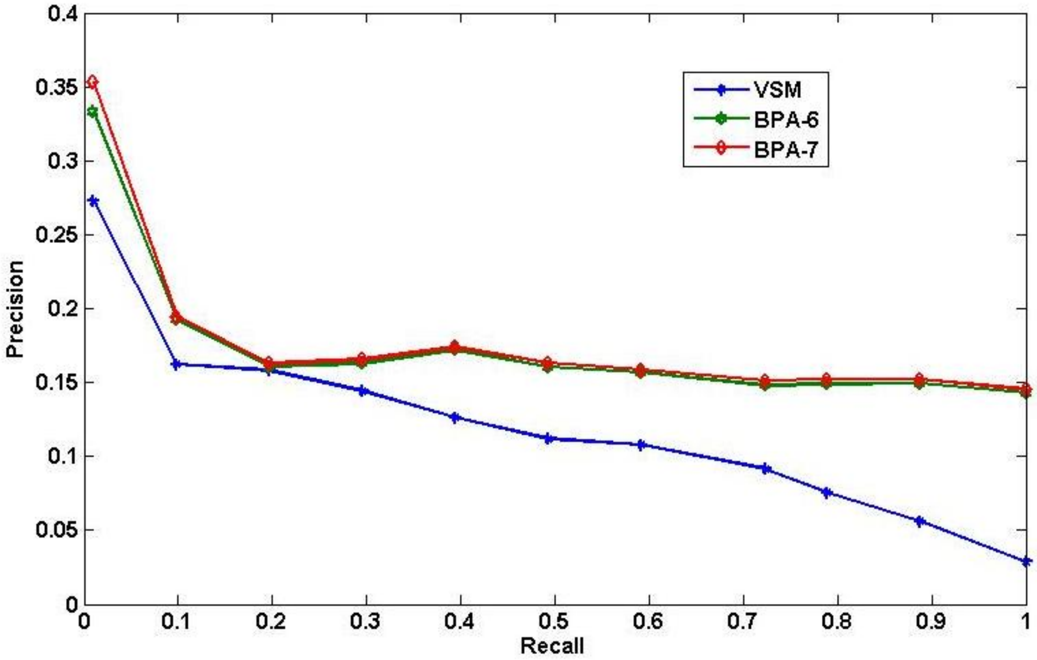

Fig. 2 Comparison between the proposed BPA-7 and BPA-6 and VSM model

5. CONCLUSION

Information retrieval model based on the Dempster-Shafer theory of evidence is used. First VSM and Basic Probability Assignments of DST have been implemented. In DST-BPA, By taking term pairs as pieces of evidence for the relevance between a document and a query and combining all the evidences, the mass function of the relevance is calculated and used to rank the documents. Collection information is also considered as one of the source of evidences along with term frequency and inverse term frequency. Dempster's combination rule is applied to combine all these sources of evidences to rank the retrieved documents. DST-BPA-6 and BPA-7 both outperform the VSM. It is also found that BPA-7 outperforms BPA-6.

Implementation of basic probability assignment which leads to improved precision and recall value to rank the documents is the future area of work.

6. REFERENCES

[1] Salton, G., Wong, A., Yang, C.S., “A vector space model for automatic indexing”. Communications of the ACM. vol-18, 1975.

[2] Theophylactou M, Lalmas M. A , “Dempster-Shafer belief model for document retrieval using noun phrases”, In: Proceeding of BCS Information Retrieval Colloquium.Grenoble, France, 1998 pp 213-228.

[3] Shi L, Nie J Y, Cao G., “Relating dependent indexes using Dempster-Shafer theory”, In: Proceedings of the 17th ACM Conference on Information and Knowledge Management.Napa Valley, California, USA, 2008: pp 429-438.

[4] Lalmas M, Moutoginni E.,“A Dempster-Shafer indexing for the focused retrieval of a hierarchically structured document space: Implementation and experiments on a web museum collection”, In: Proceedings of RIAO, 6th Conference on Content-Based Multimedia Information Access.College de France, France, 2000 pp 53-95.

[5] Shi C, Zhang J, Deng B., “A new document retrieval model using Dempster-Shafer theory of evidence”, In: Proceedings of the IEICE General Conference. Nanjing, China, 2008:pp 746-749.

[6] Jiuling Zhang, Beixing Deng, Xing Li,“Using the Dempster-Shafer Theory of Evidence to Rank Documents”, IEEE computer science TSINGHUA SCIENCE AND TECHNOLOGY pp 241-247 Volume 17, Number 3, June 2012.

[7] Hazra Imran , Aditi Sharan, “A Framework for Efficient Document Ranking Using Order and Non Order Based Fitness Function”, IMCES 2010,Hong Cong.

[8] Ian Ruthven and Mounia Lalmas, “Representing and retrieving structured documents using the Dempster-Shafer theory of evidence: modelling and evaluation”, Journal of Documentation, Vol. 54 , pp.529 – 565, 1998.

[9] Hatcher, Gospodnetic, McCandless, “Lucene in Action”,second edition,Manning publication-2009.

[10] Stephen Robertson ,"Threshold setting and performance optimization in adaptive filtering ", Information Retrieval vol.5, pp 239–256 (2002)