IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 2, No. 3, September 2013, pp. 136~142
ISSN: 2252-8938

Text Wrapping Approach to Natural Language Information Retrieval Using Significant Indicator

Enikuomehin AO*, Sadiku JS**
* Departement of Computer Science, Lagos State University
** Departement of Computer Science, University of Ilorin

Article Info

Article history:

Received Jan 14, 2013
Revised Mar 13, 2013
Accepted Mar 20,2013

Keyword:

Information Retrieval
Quantum Logic
Significant Indicator
Textual Wrapper

ABSTRACT

This paper continues the advancement of models proposed for Information Retrieval by understanding that, the Information Retrieval task continues to draw attention as the information repositories increase. Knowing that Natural Language presentation of user's information need help to reduce the complexity of the search process, we propose the use of a well defined Significant Indicator, which uses the relevance index of terms derived from the position of the text, to perform retrieval. This is achieved by initiating a text wrapping process such that document representation in space could algebraically be measured and assigned appropriate function as similarity ratio for Query and Document. Benchmark tools for Information Retrieval were followed and experiment performed using TREC classified data implemented with TRECEVAL shows better performance against some baseline models. The paper suggests further research in the direction of the Significant Indicator as a method for large search space reduction.

Corresponding Author:

Enikuomehin AO,
Departement of Computer Science,
Lagos State University,
Badagry Expressway, Lagos, Nigeria.
Email: toyin@lasunigeria.org

1. INTRODUCTION

Information Retrieval can broadly be classified into two processes: Indexing and Retrieval. Indexing is concerned mainly with collection, parsing and majorly availability of documents while Retrieval involves the process of "bringing out" the required documents. A generalized problem for all computer system users is the problem of searching for information. This problem reoccurs with more people facing it by the day. This problem is due to one major reason, the continuous increment of the information available on the World Wide Web (www). Internet services is becoming more available and the cost of acquisition is dwelling by the day thus making it easy for any interested person to add to the already over blown information servers of the internet.

The internet began as a tool for surfing the web, and then comes the ability to search. More recently, an extensive part, which is leaping has been introduced. Leaping combines the processes of surfing and searching. The aim of an Information Retrieval (IR) model is to improve the leaping time. The continuous growth of the internet and World Wide Web has tremendously contributed to research in the domain of I.R.; unfortunately, users have not been able to adapt to this exponential growth. This has obviously led to the entanglement of information thereby making it so complex to use. Information Retrieval is aimed at solving problem such that the most appropriate documents are retrieved by a given user with such results satisfying the user's information need. The word "document" is used as a general term that could also include non-textual information, such as multimedia objects etc.

A major concern in the IR framework is the user's inability to formulate their information need. Belkins in 2005 [1] based on the idea of Parsley [2] introduced the concept of ASK (Anomalous State of Knowledge) where users do not have the ability to appropriately represent their information need. The ASK concept involves the consideration of the cognition of the searcher and a method of solving this, is by considering retrieval of Natural Language Text. Natural Languages are languages used by human. Human believe that using their natural language will enable them express their information need more appropriately.

Thus, it becomes necessary to consider appropriate models for Natural Language Information Retrieval. Many models on IR have been proposed and used; however, there are three well known formal models [3], namely VECTOR SPACE MODEL, THE PROBALISTIC MODEL AND THE LOGICAL MODEL. They evolve chronologically in that order, each has its strengths and its weaknesses; for example, the vector space model is good at representing a notion of proximity between information objects (documents, terms,...); the probabilistic model is good at representing the uncertainty involved in estimating the degree of relevance of an object to an information need; and the logical model is good at representing the weight of evidence derived from viewing retrieval as a process of (plausible) inference. However, none is able to encompass all these strengths effectively. Obviously, there is a need to represent these capabilities in a single framework. The Vector Space Model has many attributes that exceeds the performance of others but fails in three major areas i.e Inability to handle large corpus (Corpus increases continuously), Ignores semantics and been unable to handle counterfactual sentences. As a result of this, Quantum logic is considered to be of importance in this regard.

Mathematical formalism of quantum theory has been shown to be of importance in the last two decades in areas other than physics. This is due to the flexibility and completeness of the quantum structures (vector spaces, inner products, quantum probability, quantum logic connectives, etc).

In this paper, a single mathematical framework is considered for logical, probabilistic and vector aspects in IR in such a way that Natural Language text can be handled. A natural language search engine would find targeted answers to user questions (as opposed to keyword search). For example, when confronted with a question such as "which Nigeria State has the highest terrorist attack"? Conventional search engines ignore the question and instead search for the keywords 'state, terrorist and attack'. Natural language search, on the other hand, attempts to use natural language processing to understand the nature of the question and then to search and return a subset of the repositories such as web databases that contains the answer to the question. The result will be more credible and of higher relevance than results from a keyword search engine. It considers the present state of the user and its effect on the search result. It must always be recalled that the usability of any information system is dependent on the degree of satisfaction of the user. The known limitations of Logic and Uncertainty in Information Retrieval still exist largely and therefore justifies the need to formulate a proper model for IR that will be based on logic, a scheme that is considered to handle ambiguity and uncertainty [3], [4]. Vector space retrieval has proven to be a substitute where the requisite behavioral data to support ranking algorithms is not present. Vector space model has been widely used in systems like Apache Lucene. Vector space model retrieval methods are potentially richer and more sensitive than the key words matching. It is also assumed that the model is qualitative in the sense that it also represents the user query as vectors in space compared with other models. In the light of this, the study examines the useful features of this model needed to implement a complete framework for Information Retrieval. A way to solve this problem is by extending the defined relationship between the algebraic formation of quantum theory and that of Information retrieval.

The Vector Space Model, as shown, is still the most popular and most acceptable model used in document retrieval in particular and objects retrieval in general but its problem of poor performance when confronted with long and large documents has not been solved because they have poor similarity due to small scalar product and a large dimensionality [14]. Another problem in this direction is the lost of similarity due to semantics within documents with similar context and different vocabularies as they are always treated as not similar. This is because, in vector space model, it is not possible to include term dependencies into the model. All the existing models lack a unified theoretical framework to address the various challenges identified in the IR domain. Thus, on application to large collection, the retrieved results are not usually the best obtainable relevant results. The researcher is hereby interested in filling all these gaps among others by proposing a more efficient IR model using theories experimented in Quantum logic.

2. EARLIER APPROACHES

The quantum based IR framework relies on a willing dimensional representation of documents (subspace) and queries (densities). However, multi dimensional representations have been implicitly used in IR to handle negative feedback. It was later shown in [5] that contradicting results were obtained. Explicit multidimensionality has been proposed in [6] to randomly slit documents into two parts and use two

dimensional stereoscopic view of a document. Furthermore, [7] propose the use of vector as description for information need as a way that can solve some problems. A more explicit discussion on vector based representation is given in [8]. The approach in this work can be considered as a partial extension of this earlier work but the problem of effectiveness on large document set still exist. The quantum based probability makes use of Hilbert spaces, unit vectors and subspaces. The probability that a document is relevant to a users information need is determined by the projection of its vector representation also the corresponding to its information need (subspace). The roles of documents and the information need can be interchanged which is motivated by the proposition that information need should be presented as a dynamic component [9]. Subspaces are a core component of the generalization of the probabilistic framework brought by quantum physics [10], which enables us to combine both geometry and probabilities [11]. Sophisticated document representations have already been explored. The proposal of Melluci that subspaces can be used was trashed [12].

The main criticism for the vector space model is that it provides no formal framework for the representation, making the study of representation inherently separate from the relevance estimation. The separation of the relevance function from the weighting of terms has the advantage of being flexible, but makes it very difficult to study the interaction of representation and relevance measurement. The semantics of a similarity/relevance function is highly dependent on the actual representation (i.e., term weights) of the query and the document. As a result, the study of representation in the vector space model has been so far largely, heuristic [13].

3. JUSTIFICATION

The quantum world defies the rules of ordinary logic. Particles routinely occupy two or more places at the same time and don't even have well-defined properties until they are measured. It's all strange, yet true - quantum theory is the most accurate scientific theory ever tested and its mathematics is perfectly suited to the weirdness of the atomic world [15]. Human thinking, as many of us know, often fails to respect the principles of classical logic. The study makes systematic errors when reasoning with probabilities, for example. Physicist Diederik Aerts of the Free University of Brussels, Belgium, has shown that these errors actually make sense within a wider logic based on quantum mathematics. The same logic also seems to fit naturally with how people link concepts together, often on the basis of loose associations and blurred boundaries. That means **search algorithms based on quantum logic could uncover meanings in masses of text more efficiently than classical algorithms.** Classical logic has problems in this situation, because in set-theory if a is an element of the union $A \cup B$, it follows that at least one of the statements $a \in A$, $a \in B$ must hold — for example, where a represents the state of an electron which has passed through the two slits A and B then one of the statements "a passed through A" or "a passed through B" must hold. Quantum logic solves this problem by describing the outcomes A and B not as arbitrary sets, but as subspaces of a vector space. Their disjunction is then their vectors sum A+B, which is strictly larger than their set union $A \cup B$ unless $A \subseteq B$ or $B \subseteq A$. Since there are many points $a \in A + B$ which are neither in A nor in B, the question "which slit did the electron a go through?" ceases to apply.[16] Putnam (1976) contends that the differences between quantum logic and classical logic can account for all of the apparent 'difficulties' of quantum mechanics, and the structure of quantum logic itself is quite simple and is arrived at, precisely by replacing the notions of sets and subsets with those of vector spaces and subspaces. This leads us to consider the collection $L(V)$ of subspaces of vector space V, which is a partially ordered set under the inclusion relation, so that an event A implies an event B precisely when $A \leq B$.

Quantum logic differs from classical Boolean logic in (at least) two well-known properties: quantum logic is neither distributive nor commutative. The distributive law

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

is responsible for the question "which slit did the electron pass through", and so (as described above), quantum logic avoids this issue by avoiding the assumption that the electron must have passed entirely through either one of the slits.

4. MODEL AND EVALUATION

In the evaluation of the proposed retrieval methodology, two main parameters were used: the number of term and the window size also referred to as width. For the purpose of the test, a maximum size was fixed but in practice, it is believed to have some performance ability on undefined window size. The number of terms is assumed to be greater than the number of query terms. This is used to generate a ranking procedure in the following manner:

1. Pre assignment of weight to query terms is carried out so that non common terms including stop words will be considered in weight generation. This also affects the definition of the window size.
2. A weighted probability distribution of the text shape is considered in terms of the distance at which the query terms appear. This enables the shape to be obtained over a given document distance. The weighted distance will be used for building the score for each term.
3. The high scoring terms are selected and normalized with their shape such that the highest score is one for terms and the shape sums to one.
4. Score are computed by applying the Significance Indicator to each document. This is done by assigning a number to each position and generating a score from all textual shape. This score is added to overall score and a power of the shape is chosen to allow for weight overlap.

4.1. Significance Indicator

The Significance Indicator (SI) is used to complete this work as a basis of setting up an evaluation process for it. The SI is an operator that scores a document according to the degree to which it is significant. More precisely, The Significant Indicator shows the degree of significance between documents in a collection and uses properties in linear algebra and the uncertainty condition to generate the Indicator index. This is developed from the idea of uncertain conditional as an operator, an Indicator that checks if a document is related to another document which is then used to formulate some query terms representing the documents. The Significant Indicator is built such that it conforms to some algebraic properties defined for document representation.

The idea of the SI is based on the consideration of the linear space of the Text Wrapper (TW), the linear combination of the space which is a dual space to that of documents: This is a new idea worth exploring.

4.2. Discussion

The textual wrapper reduces the number of NL texts in any document and therefore one could relate to the Shannon information theory [17] as it infers a complete measure for information over a space. For any position in a word space, it is important to know the text that occupies it. If this is unknown, then this will constitute to a flat probability throughout the vocabulary, where probability is

$$\frac{1}{S_V}\left(S_v = size\ of\ \ vocabulary\right) \tag{1}$$

If we know what term occupies a given position, then we have a best situation that will correspond to a distribution where one term has a probability of 1 and the rest have probability of 0. We can generate a scale for positions in an information which can be contained by the text wrapping process will have 1 units and uncontained text will have 0 units, this can be given as:

$$K_{position\ x} = \frac{\log\left(S_V\right) + \sum_t P\left(t \in postion\ x\right)\log p\left(p\left(t \in x\right)\right)}{\log\left(S_V\right)} \tag{2}$$

If the terms in such document are independent, the total information contained in the document will exactly be the length when all terms are contained. This will change within the textual wrapper has acted on it. Some of the positions will contain terms that are unknown or undetermined, this is the only conclusion for such position is that they have it is that a central term of the TW. Then, the amount of information would have changed, we have:

$$K_{unwrapped\ position} = \frac{\log\left(S_V\right) + \left(S_V - 1\right)_{\frac{1}{S_V - 1}}\log\left(_{\frac{1}{S_V - 1}}\right)}{\log\left(S_V\right)} \qquad = 1 - \frac{\log\left(S_V - 1\right)}{\log\left(S_V\right)} \tag{3}$$

This gives the opportunity to formulate a system for the wrapper terms (that is, position that are considered known after the TW has been applied), such that we have, the total information in the contained position to be

$$K(W(t,w)D) = N_u + (L_D - N_u)\left(1 - \frac{\log(S_V - 1)}{\log(S_V)}\right) = N_u\left(\frac{\log(S_V - 1)}{\log(S_V)}\right) + (L_D - N_u)\left(1 - \frac{\log(S_V - 1)}{\log(S_V)}\right) \quad (4)$$

Where L_D is the length of the document, S_V is the size of the vocabulary and N_u is the number wrapped text. $\frac{\log(S_V - 1)}{\log(S_V)}$ will be close to 1 when considering large vocabularies, that is for large S_V. Example, for a vocabulary of 10 terms; it will be 0.0458, for 20 terms; 0.01712, for 100 terms it will be 2.10 x 10^{-3} etc. Thus this count of information is equivalent to counting of the wrapped text terms. The above model suites our definition for a representation of relevant text.
[18], [19]. It has been shown that the distance between terms can be used as a criterion for selecting keywords. This is normalized as a sequence of terms and a frequency can be given as
This can be applied to any given text to generate the keywords.

Similarly, using a given set of frequencies to represent a document can bring large number of documents. It's large that the vector of their term frequencies is used to represent them and given as:

$$N_{Sequence} = \frac{(\Sigma_t N_t)!}{(\Pi_t (N_t!))} \quad (5)$$

This can be verified by using a document collected in TREC, labelled AP880121001 which has 249 distinct terms and a sequence of 383 tokens, distributed as:

Table 1. TREC AP880121001 Collection

N_t	25	15	8	7	6	5	4	3	2	1
Terms with N_t	1	1	2	1	3	6	5	9	38	149

Using (AP88), we find if vector for the document would be generated with high number of possible sequence:

$$N_{Sequence} = \frac{382!!}{(25)!(15)!(8!)^2(7!)^2(6!)^3(4!)^5(3!)^9(2!)^{38}} \approx 4.47 \times 10^{738} \quad (6)$$

These sequences will not comply with the syntax structure of natural language. This it will be very difficult to generate a fraction or part of the sequence that makes sense. We use a standard measure to test this. (the mean average precision). The mean average precision is the mean of the average precision score for each query. Generally it is given as

$$MAP = \sum_{q=1}^{Q} aver\ p(q)/Q \quad (7)$$

Where Q is the number of measure of a ranked retrieval run. It is the mean of the precision score after each relevant retrieval. The values for this are the mean of the individual average precision score. It has the traditional recall and precision aspects which is sensitive to the entire ranking. The MAP has a lesser stressful interpretation than other frequently used measure. The MAP is computed by scanning document position and computes the precision value for the document from top to each occurrence of a relevant document. The retrieved documents can be considered as a set

$$\{R_i\} = \sum_{i=1}^{NR} i/RiNR \quad (8)$$

Where NR is the total number of relevant document. MAP is essential to this work because it assumes that non-assessed documents are irrelevant. This is a reason why other measure can be considered in addition to the MAP. One can consider other measures such as Jelinek Mercer which involves the linear interpolation of the Machine Language model with the collection models. It is commonly used effectively in

smoothing. More confidently, is the Bpref measure which is close to the MAP but counts all assessed documents. For the scope of this work, only MAP is used to test this process.

4.3. Data Set

The data used in this thesis is the TREC I data which has been shown to be a very successful and useful collection. The TREC I collection used \are:

(a) WSJ87-89
The copyrighted stories from the WSJ for the 87-88 are used. They also include those of 1989. This is available from Dow Jones information services.
(b) AP 89 (AP 89 includes the copyright newswire collected by AT & T Bell from 1989).

5. FINDINGS AND RESULTS

The formalism described earlier in the thesis and outlined in the behavior of the application of the Q/D significance indicator (SI) has a performance level that are acceptable in the same vain as other methods such as (TF, TFIDF, BM2X, and languages model). The advantage of this model is in its ability to handle long sentences and large corpus. The results are presented in a table form described below;

Table 2. Performance of SI against other known Models

MODELS	4	10	avg	4	10	avg
MAP for LMDP		14.33%			13.09%	
MAP for BM 25		18.44%			16.79%	
MAP for TFIDF		14.54%			13.99%	
MAP for TF		9.88%			7.16%	
MAP for SI (10 terms)	13.46	13.65%	14.15%	13.67%	13.54%	12.80%
MPA for SI (15 terms)	13.66	13.62%	13.75%	24.20%	13.85%	13.04%

The covering width is the minimum width of a given text wrapper of the text term that are wrapped in a whole document and the percentage of covering (%) is the covering width as a percentage of the length of the document T. Application to the TREC collections:

Table 3. TREC AP and WSJ collection details

Collection	AP 89	WSJ8789
No of document (1000)	84.68	98.73
No of term (x1000)	207.62	169.34
No of token (x100000)	41.80	43.68
Avg length	493.66	442.44
Avg covering width	432.27	681.75
Avg covering %	72.71	70.33
Avg doc with term	104.289	127.79
Avg occurrence	1.38	1.46

A major motivation for this work is on issues involving large NL terms. We therefore consider the limit in relation to number of terms; since we used a fixed number of term for the construction of the SI, we need to evaluate the effect of this number.
This test is carried out using Treceval.
Generally, the shapes are generated following the scoring of each textual relation (distance) from O to a maximum. In this, the identity of the term is the distance is considered since query terms one already assigned a score that will be counted and considered by the wrapper every time the query term appear in a considered distance.

6. CONCLUSION

Many attempts were made to limit the scope of this work to the defined objective; however, it is necessary to establish the basic concepts defined earlier. The mathematical formulation became a basis to generate the necessary framework for the retrieval model. In that framework, we consider the quantum theory with its logic as a basis for generating a relation between text terms given a document space. This concept is

similar to measurement and observation in the quantum sense. This helps to realize that once the defined textual wrapper operates on a document, the content of that document is preserved in subsequent operation. This paper is a theoretical approach for solving some known IR problem essentially in areas where the fundamental baselines models have not yielded maximum expected result. The aim of every IR researcher is to consider the same function of the user and this is why the study taken this to the National language level. The paper's introduction of the Significant Indicator is considered fundamental and a new starting point for IR research, it is also important to mention that efforts were made to associate the concept of this research with some existing theories such as the properties of a Boolean set where intersection as example, is presented or TW meet etc.

REFERENCES

[1] NJ Belkin. Anomalous State of Knowledge. IN: KE Fisher, S Erdelez, & EF McKechnie (Eds.). Theories of information behavior: A researchers' guide. *Medford, NJ: Information Today*. 2005: 44-48.

[2] WJ Paisley, EB Parker. Information Retrieval as a Receiver-Controlled Communication System. IN*: Education for Information Science* London: McMillan. 1965: 23-31.

[3] PD Bruza, F Crestani, and M Lalmas. *Second Workshop on Logical and Uncertainty Models for Information Systems*. In Proceedings of DEXA 2000. IEEE Press, Greenwich, London, UK. 2000.

[4] F Crestani, M Lalmas, and CJ van Rijsbergen. *Information Retrieval: Uncertainty and Logics*. Kluwer Academic Publisher, Norwell, MA, USA, 2000.

[5] G Salton and M McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, NewYork. 1983.

[6] C Zuccon, L Azzopadi and CJ Van Rijsbergen. Semantic Spaces: Measuring the Distance between Different Subspaces. *In QI*. 2009: 141-123.

[7] L Che, J Zen and N Tokud. A "stereo" document representation for textual information *JA SIST* 5. 2006.

[8] MW Berry, Z Drmac, and ER Jessup. Matrices, Vector Spaces,and Information Retrieval. *SIAM Review*. 1999; 41(2): 335-362.

[9] CD Manning, P Raghavan, & H Sch¨utze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK. 2008.

[10] MA Nelsen and IL Chuang. *Quantum computation and Quantum information*. Cambridge, UK: Cambridge University Press. 2000.

[11] CJ Van Rijsbergen. *The geometry of information retrieval*. Cambridge universal press. 2004.

[12] P Bruza, D Song. Towards Content Sensitive Information Inference. *Journal of American society for information Science and Technology*. 2003; 54(3): 321-334.

[13] M Melucci. A basis for information retrieval in context. *ACM TOIS*. 2008; 26(3).

[14] D Widdows. Semantic vector products: Some initial investigations. *Second aaai symposium on quantum interaction, oxford, AAAI*. 2008: 1–12.

[15] A Huertas-Rosero, L Azzopardi, C Van Rijsbergen. *Characterizing through Erasing: A theoretical framework for representing documents inspired by quantum theory in PD Bruza*. W Lawless CJ VR, ed: Proc. 2nd AAAI Quantum interaction symposium, Oxford, U.K. College publication. 2008: 160-163.

[16] GM D'Ariano, R Demkowicz-Dobrzanski, P Perinotti, and MF Sacchi. Quantum-state decorrelation. *Phys. Rev. A 77*, 032344. 2008.

[17] Putnam H. *The logic of quantum mechanics. In Mathematics, Matter and Method*. Cambridge University Press. 1976: 174–197.

[18] C Shannon, W Weaver. A mathematical theory of communication. *The Bell System Technical Journal*. 1928: 379–423.

[19] P Carpena, P Bernaola-Galván, M Hackenberg, AV Coronado JL. Level. 2009.

[20] M Ortuño, P Carpena, P Bernaola-Galván, E Muñoz, AM Somoza. Keyword Detection in Natural Languages and DNA. *Europhysic Letters*. 2002; 57: 759-764.