Context Vector Machine for Information Retrieval

V. Khatavkar[1] and P. Kulkarni[2]

[1]Department of Computer Engineering and Information Technology, College of Engineering, Pune, MS, India.
[2]Department of Computer Engineering and Information Technology, College of Engineering, Pune, MS, India. {vkk.comp@coep.ac.in, paragindia@gmail.com}

Abstract. Information Retrieval (IR) is the need of todays digital world. It is important for a user to get appropriate document from the sets of documents. In order to do so, researchers are working in the domain of IR mainly on the concepts like Document Clustering and classification, Ontology of document, Thematic of document, Concept of document, Context of document, etc. The research work proposes necessity of Context Vector Machine.

Keywords: ***Information Retrieval; Context; Context Vector; Document Clustering; Document Classification***

1 Introduction

There are various vector space models proposed by researcher like Salton et. al. [1, 2, 3]. The vector space model is chosen in order to improve recall and precision of the document retrieval system. In vector space model, the documents are the represented in the vector space and then the vectors are computed, which in turn can give the user useful information with improved recall and precision. After the user gets the useful information, it is validated by Kappa Statistics, which is given as:

$$K = P_A - P_0 / 1 - P_0 \quad (1)$$

Where, P_A is the probability that two reviewers agree in practice, and P_0 is the probability that they would agree solely by chance. K is always between 0 and 1, with 0 indicating no better agreement than expected by chance and 1 indicating perfect agreement.
Performance is calculated based on recall, precision and accuracy using F-measure. The various formulae used are given below:

$$\text{Recall} = \text{Number of correct positive predictions} / \text{Number of positive examples} \quad (2)$$

$$\text{Precision} = \text{Number of correct positive predictions} / \text{Number of positive predictions} \quad (3)$$

$$\text{F-measure} = 2 * \text{Precision} - \text{Recall} / \text{Precision} + \text{Recall} \quad (4)$$

For many Information Retrieval (IR) systems above formulae are used frequently. The coming sections, explains the necessity of context vector space models, the current state of art in vector space model and the necessity of context vector machine. The paper concludes with the proposed system which uses context vector machine.
Graphs are very useful tools to represent documents. Sonawane et. al. in [5] have surveyed various techniques which are used to analyze text documents using graphs. The relation can be co-occurrence, grammatical or conceptual.

2 Necessity of Context Vector Space

Traditional approaches use concept of “word matches" for document retrieval. The results given by the system, developed using “word matches", consists of large number of documents which is not effective from the user perspective. In order to reduce the number of documents in result, more number of words are to be given in

B. Iyer, S. Nalbalwar and R. Pawade (Eds.)
ICCASP/ICMMD-2016. Advances in Intelligent Systems Research.
Vol. 137, Pp. 375-379.
© 2017- The authors. Published by Atlantis Press
This is an open access article under the CC BY-NC license (http://creativecommons.org/licens)es/by-nc/4.0/).

query. But this increases the risk for user, since the words in query given by user might not be of interest in the document. Main drawback of this system is low recall rate.
In order to improve precision and recall, vector space model is proposed. Salton et. al. in their work [1, 2, 3] , represented document as vector. The vector space consisted of a number of coordinates that was equal to the number of unique terms in the system (the vocabulary). The direction of the document vector was determined by counting the number of terms in the document, then applying an appropriate term weighting and normalization [4, 6] to these counts. This approach produces what we will refer to as a "term orthogonal" space. That is, each term in the vocabulary is a coordinate and all coordinates are, by definition, orthogonal.

3 Vector Space Models

There are various models proposed for Information Retrieval (IR) namely,

1. Boolean Retrieval: It uses Boolean operators for IR. Fuzzy set theory is typically used in this model.
2. Word Association: Concepts like word stemming, dictionary and relevance feedback are used in queries.
3. Document Representation: Concepts like Term co-occurrence and Term Frequency are used.
4. Vector Space Model: Concept like Document Indexing, Term Weighting and Similarity Coefficients are used.
5. Probabilistic Retrieval: Probabilistic models are developed.
6. Latent Semantic Indexing: IR models suffer from two well known language related problems called synonymy and polysemy. So while developing a model, the terms meaning i.e. semantic is taken into consideration.
7. Document Classification: Unsupervised classification (clustering) and supervised classification are used. In unsupervised classification, hierarchical clustering shows better results than other clustering methods while in supervised classification, bayes classifier and subspace method outperform other methods of classification.

According to Salton et. al. [7] the concept of context vector is by representing documents in terms of vector as:

$$D_i = (D_{i1}, D_{i2}, \ldots . D \ ) \quad (5)$$

where d_{ij} represents weight of the j^{th} term. If the index of two documents is given then, similarity coefficient between them can be $s(D_i;D_j)$ which reflects the degree of similarity in the corresponding terms and term weights. Such similarity coefficients can be calculated using methods like cosine similarity and TF-IDF.

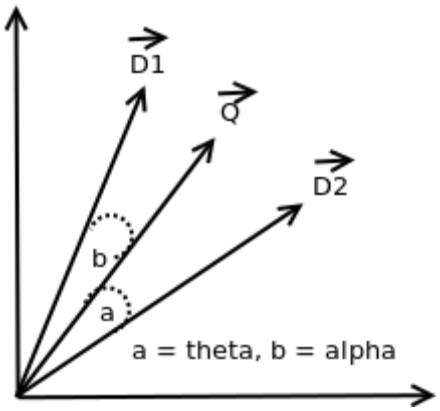

Fig.1. Vector Space Model

Cosine similarity for a query, q, for vectors shown in shown in Figure 1 can be given as:

$$cos(\theta) = \frac{d_2 * q}{\|d_2\| * \|q\|} \quad (6)$$

where, $\|q\|$ is norm vector given as :

$$\|q\| = \sqrt{\sum_{i=0}^{n} q_i^2} \quad (7)$$

If Cosine Similarity is zero then there is no match between document d and query q. Such document and query is said to be orthogonal. The second method for calculating similarity coefficient is TF-IDF i.e. Term Frequency and Inverse Document Frequency. In this model, term-specific weight is used. The weight vector for document d is $v_d = \left[w_{1,d}, w_{2,d}, \ldots w_{N,d}\right]^2$, where ,

$$w_{t,d} = tf_{t,d} * log \frac{|D|}{\{d' \epsilon D \vee t \epsilon d'\}} \quad (8)$$

Here, $tf_{t,d}$ is is Term Frequency of term t in document d and $log \frac{|D|}{\{d' \epsilon D \vee t \epsilon d'\}}$ is Inverse Document Frequency. $|D|$ is the total number of documents in the document set while $d' \epsilon D \vee t \epsilon d'$ is the number of documents containing the term t.

Vector Space Model is advantageous since it is based on simple linear algebra.

But for long documents it has poor similarity values. It is semantic sensitive. In order to handle long documents/ high dimensional document sets, Latent Semantic Indexing (LSI) is used. The basic LSI model extracts information in vector form by identifying the dependencies between terms and documents.

This approach is based on constructing and processing a term-by-document matrix. In LSI, the resulting dimension of the vector model is determined via singular-value decomposition of the term-by-document matrix and, in general, will be signifcantly smaller than either the term or document count.

As a consequence, the vectors that represent words cannot be orthogonal. As such, the LSI approach encodes a form of similarity of usage, but at a low resolution, term-by-document level.

The key innovations behind the MatchPlus approach are motivated by the desire to exploit neural network learning techniques to discover similarity of usage at the word level, in a language-independent manner, without the need for external dictionaries, thesauri or semantic.

4 Context Vector Machine

Gull et. al. in their work [8] uses concept of Automatic Web Page Categorization by Link and Context Analysis. Haribhakta et. al. worked on context of document [9] . According to [9] , "Context is the theme of the document". Caid et. al. in their research work [10, 11, 12] came up with the concept and implementation of "Word Vector". Parag Kulkarni [13] states that he has worked on various research projects and worked for multi-perspective and context based decision making. From this, he comes up with an invention of Context Vector Machine (CVM); which according to him can solve problems in document mining and image processing.

5 Necessity of Context Vector Machine

Sanjeev et. al. [14] in their patent invented a method of doing business with various documents namely papers, images and electronic documents. These can be processed and analyzed using computers. They identified a major problem in manual organization of documents is time consumed for collation of papers documents and extraction of information. This highly affects the decision making process in business. Thus there was a need of automation in this process. State of art in the process of fast business decision making automation are as mentioned below :

1. Some of them are completely dependent on image processing techniques.
2. Some of them use Bayesian and/or Support Vector Machine (SVM) algorithms for text classification which are nothing but usage of simple keyword search techniques.
3. Some of them rely on document boundary detection methods for text and image classification techniques.

Though today many documents are in the form of electronic format like pdf and Optical Character Recognition (OCR). The challenge is that for low image quality the systems performs very poor. If there is a large sets of e-documents to be processed, they came across following problems :

- The systems are limited to document boundary detection, document classification and text categorization.
- They offer document collation with separation of very similar documents.
- They also do not offer domain-sensitive scrubbing of extracted information.

Techniques based on the current methods of Information Retrieval without considering context and keyword based classification cannot offer the consistent extraction of information from documents for automated decision making. If technique of similarity is used then similarity among documents may lead to misclassification when using pattern-based classification. For structured data if extraction is done using template-based matching generally failson even slight shift of images and if done with rules- based templates can return false results if there are significant variations of the document. The system can't handle documents (structured and unstructured documents) equally, efficiently and reliably.

6 Proposed System

Context Vector Machine (CVM) based System. In preferred embodiments of the instant invention, the collation process is based on incremental learning and various Artificial Intelligence (AI) based techniques, which may include one or more of the following, such as:

1. The Location Diagram and Feature Vector-based feature extraction and page mapping,
2. Support Vector Machine (SVM) and Natural Language Processing (NLP),
3. An intelligent filter technique taking advantage of header and footer based information,
4. Collation by finding common threads within or between pages, documents, or sets of documents,
5. Finding disagreements based on affinities,
6. Inference-based mapping,
7. Feature based discontinuity detection and collation, as well as human collaboration.

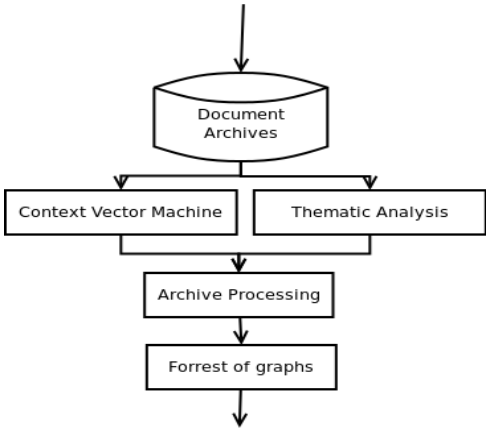

Fig.2. Proposed System

The proposed system which according to literature findings and our conclusion should be able to classifying at least some of the discrete document pages using the sets of text-based information, wherein multiple classification engines are employed and classification is based on a consensus of the classification engines.
The proposed system will focus on Support Vector Machine (SVM) and Natural Language Processing (NLP).
When a look is taken on the literature survey done by researches in the domain of document analysis, specifically in the area of defining context, context vectors, Thematic, SVM and CVM, it motivates the researcher to work on linking above said areas. This research work will be an attempt to do same. The proposed system is shown in Figure 2. The proposed system will work in following order:

- The proposed system will take document archive as input.
- It will then process the archives with Context Vector Machine (CVM) as well as perform Thematic Analysis.
- Then it will perform processing on the processed archives to generate graph for each document.
- It will form forest of graphs.
- The user may take a graph/s; which is our intended output.

The document archive can be a standard document dataset or text dataset or real documents. In proposed system the output given by CVM will play a crucial role. The output of CVM and Thematic analysis will be processed on the archives given as input and will generate forest of graphs.
The forest of graphs can be directed or undirected; hierarchical or single level.
Graphs can be dependent on each other or independent of each other. The user can get his output as either a graph or set of graphs which are related to each other. They can be undirected, directed and/or hierarchical graphs. The nodes will represent topic in a document while the links will represent relation between the topics.

7 Conclusion

Context vector machine (CVM) can be used for automatic information retrieval in e-documents. The proposed system focuses on design of CVM along with the thematic analysis of document. The proposed system will be helpful to cluster and/or classify document/s with respect to theme of a document/s which is nothing of context of a document/s. In the proposed system the context of document/s will be converted into context vector space where different formulas will be applied to get the desired output. The clustered and/or classified set of

document/s can be again processed and converted into graphs. The various graphical representation techniques can be applied.

References

[1] Salton, G. (ed.), "The SMART Retrieval System - Experiments in Automatic. Document Processing", Prentice-Hall, 1971.
[2] Salton, G., "Another Look at Automatic Text Retrieval Systems", Communications of the ACM, Vol. 20, 1986.
[3] Salton, G., "Automatic Text Processing", Addison-Wesley, 1989.
[4] Salton, G. Buckley, C., "Term Weighting Approaches in Automatic Text Retrieval", Information Processing and Management, Vol. 24, No. 5, 1988.
[5] S. S. Sonawane And Dr. P. A. Kulkarni,"Graph Based Representation And Analysis Of Text Document: A Survey Of Techniques" , International Journal Of Computer Applications, 2014.
[6] Buckley, C., Allan, J., Salton, G., "Automatic Routing and Ad-hoc Retrieval Using SMART: TREC-2", in Proceedings TREC-2 Conference, D. Harman, ed, Gaithersburg, MD. Aug. 1993.
[7] Salton, G. and Wong, A. and Yang, C. S.,"A Vector Space Model for Automatic Indexing", Communications of the ACM,Volume 18 Issue 11, 1975.
[8] F. S. Giuseppe Attardi, Antonio Gull, Automatic web page categorization by link and context analysis, 2000.
[9] Y.V. Haribhakta, Dr. Parag Kulkarni, "Learning Context For Text Categorization", International Journal Of Data Mining & Knowledge Management Process , 2011.
[10] W R Caid, P Oing, "System And Method Of Context Vector Generation And Retrieval", US Patent 5,619,709, 1997
[11] W R Caid, J L Carleton, P Oing, D J Sudbeck, "Context Vector Generation and Retrieval" , US Patent 7,251,637, 2007.
[12] William R. Caid And Joel L. Carleton, "Context Vector-Based Text Retrieval", A Fair Isaac White Paper, August 2003.
[13] Parag Kulkarni, "Knowledge Management And New Paradigm Of Advanced Machine Learning", Habilitation Thesis, Faculty Of Informatics, Masaryk University, Czech Republic, 2011.
[14] Sanjeev Malaney, Parag Kulkarni, Krishnawami Viswanathan and Vikram Malaney, "Business method using the automated processing of paper and unstructured electronic documents.", US 20070118391.