I.J.Modern Education and Computer Science, **2012, 4, 36-42**
Published Online May 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2012.04.05

A Technique to Choose the Proper Vector Space Models of Semantics in Case of Automatic Text Categorization

Sukanya Ray
Amity School Of Engineering & Technology, Amity University, Noida (U.P.), India
sukanyaray007@gmail.com

Nidhi Chandra
Amity School Of Engineering & Technology, Amity University, Noida (U.P.), India
nsrivastava5@amity.edu

***Abstract*—**vides a proper solution to this limitation. There are broadly three main categories of Vector Space Model: term-document, word-content and pair-pattern matrices. The main aim of this paper is to discuss broadly the three main categories of VSM for semantic analysis of texts and make proper selection for automatic categorizing. The scenario taken up here is categorization of research papers for organizing a national or an international conference based on the proposed methodology. Computers do not understand human language and this makes it difficult when human wants the computer to do some specific task like categorization according to human need. Vector Space Model (VSM) for semantic analysis of texts and make proper selection of one of the three main categories for automatic categorizing of research papers for organizing a national or an international conference based on the proposed methodology.

***Index Terms*—**Vector Space Model, Term Document, Word Content, Pair Pattern

I. INTRODUCTION

Computers understand very little human language and that makes task very difficult to accomplish. It becomes very difficult for computer to accomplish what is exactly asked for by human. During the time of admission thousands of application forms are being submitted and manually categorizing it according to courses is a very difficult and time consuming task; in various multi-national companies every day people apply for various posts and here also manually checking whether the applicant matches the criteria and then categorizing manually the applications according to the posts is very difficult; same problem is there while organizing any national and international conference also. In this case the thousands of research papers are submitted daily and categorizing this into domains and that too in a very short duration is extremely difficult task.

But this automatic categorization according to the need of human is very difficult and the result is most of the times not satisfactory. Vector Space Model of semantics has provided a solution to this obstacle. Semantic means meaning of words, phrase, sentences or text in human language and the study of that meaning and here it is used in that particular sense. Vector Space Model on semantics is broadly categorized into three main categories: i) term-document matrix ii) word-content matrix and iii) pair-pattern matrix [1].

A. Term-Document Matrix:

In this method the similarity between documents is found. In case of a large collection of document there will be a large number of document vectors and it is easy and most convenient to arrange the vectors into a matrix. The row vector corresponds to the number of

individual terms in the document and the column vector corresponds to the documents. This kind of matrix is called term-document matrix. Here a bag of word hypothesis is used where for example {x, x, y, y, z, z} is a bag of word having term x, y, z. The order of the arrangement of the terms is not important as { x, x, y, y, z, z } and {z, x, z, y, y, x} are equivalent and the bag of word {x, x, x, y, z, z} can be represented with a vector A = {3, 1, 2} where the first element of A is the frequency of x in the bag, the second element is the frequency of y and the last element of A is the frequency of z. A set of bags can be represented as a matrix M, where each row m:j corresponds to a bag and each row mi: corresponds to a unique term in the document and an element m i:j is the frequency of the i th term in the j th bag. In case of information retrieval, with the bag of word hypothesis the relevance of the document to a query is estimated by taking the query and the document as a bag of word.

B. Word-Content Matrix:

In this method the similarity between words are found. In word content matrix the content may be given by a sequence of characters, words, phrases, sentences, paragraphs, texts, chapters, documents. It follows the distribution hypothesis which is that words which occur in similar contexts tend to have similar meaning. The idea that similarity between words can be found and used has found its application in various language games. Firth said “You shall know a word by the company it keeps.”

C. Pair-Pattern Matrix:

In this method the similarity of relationship between terms are found. In pair pattern matrix row vectors correspond to pair of words, for example footballer: football, driver : car and the column vector corresponds to the pattern in which the pairs occur, for example in the first one “A plays with B” and in the second one “A drives B”. Pair pattern similarity can be used to find similarity of a sentence in both active and passive form, for example “A does B” and “B is done by A” tends to find a similarity between A: B. the latent relation hypothesis used here states that pair of words that co-occur in similar patterns tend to have similar semantic relationship. This hypothesis is also the inverse of the extended distributional hypothesis which states that pattern with similar column vectors in the pair pattern matrix tends to have similarity in meaning. Based on the type of application each category can be chosen and based on the hypothesis used, the desired task can be performed.

This paper is divided into 4 sections. The second section focuses on the existing applications. The third section gives the proposed methodology to be adopted and last fourth section will give summary and direction for future research.

II. EXISTING APPLICATIONS

There are various applications which exist for each category. Based on these applications it can be chosen which category must be used for automatic categorization of documents for organizing conference. The applications of term-document matrix are: Document Retrieval: This was the first application developed using term-document matrix. The idea was if queries is submitted then arrange the documents in descending order of cosine of the angle between the query vector and the document vector [2].

Document Clustering: If the measure of the document similarity is given then the documents can be clustered into groups such that the similarity between the documents is high within a group and low across a group [3].

Document Segmentation: This application segments the main document into sections where each section focuses on each subtopic of the document [4].

Question Answering: If a simple question is asked then this application can answer it in short by searching for the answer in the corpus. This application is based

on document retrieval and passage retrieval applications [5].

Call Routing: Based on caller's spoken answer to question this application can automatically route telephone calls. If the caller's answer is ambiguous, then the system automatically generates another question for the caller to answer [6].

The applications of word-content matrix are: Word Similarity: Word similarity can be found by using term document but when the size of the document is very short, say 50 words in comparison to 150 words then word-content matrix gives better result [7].

Word Clustering: This application uses soft hierarchical clustering to row-vectors in a word-content matrix [8].

Word Sense Disambiguates: This application uses a feature vector representation in which each vector corresponds to a token of word [9].

Context-sensitive Spelling Correction: People often make mistakes with the usage of words like there, their, they're and this incorrect usage cannot be detected by dictionary spelling. Here relation between words will come into play [10].

Semantic Role Labeling: The task of this application is to label parts of a sentence according to the roles they play in a sentence, usually by matching their connection with the main verb used in the sentence [11].

Query Expansion: Often it is seen that the query submitted to web browsers do not directly match the terms in the most relevant documents. To solve this problem the query expansion is used for generating new search terms that are consistent with the intention of the original query [12].

The applications of pair-pattern matrix are: Relational Similarity: With the help of cosine of an angle between row vector, the attribution similarity can be measured by the word-content matrix. But relational similarity can be measured by the relational similarity between rows in a pair-pattern matrix [13].

Pattern Similarity: In this application instead of matching the similarity between the row vectors, the similarity between the column vectors can be matched and thus pattern similarity can be measured [14].

Relational Clustering: This application clustered word pairs and represented them as row vectors in a pair-pattern matrix [15].

Automatic Thesaurus Generation: Word-content matrix can be used for the Thesaurus generation but pair-pattern matrix gives better result as similarity between relationships is more important and useful than similarity between words for the Thesaurus generation [16].

Analogical Mapping: Generally proportional analogies have the form a:b::c:d, which means that "a is to b as c is to d". For example football: footballer :: basketball: basketball player. With the help of this application the proportional analogies can be solved by selecting the proper choice that maximizes relational similarity [17]. For example (similar (football: footballer, basketball: basketball player) has a very high value).

III. PROPOSED METHODOLOGY

From the study of the various existing applications it is clear that text categorization can be done using any one of the three categories. If the pair-pattern matrix is considered with the application as building an automatic research paper categorization for organizing a conference then finding similarity between relationships is a major area. The relationships can be like sore throat can be treated as cause and the paper can be from a medical domain but when it comes to papers from computer engineering background then finding a relational pattern between terms and also building a proper corpus for matching with the relationship found and then proper categorization of the uncategorized document based on this result is a fairly difficult task. Even if relationships are found between terms then building a suitable corpus so that

the document gets properly categorized is a major issue. So considering such difficulties the pair-pattern category is not a suitable one for automated research paper categorization for organizing a conference.

Now we can take in to consideration the word content matrix category for this application as the word content matrix depends on the similarity between the terms or it tries to find the context in which the term is being used with the help of the neighboring terms. Let us consider that two research papers are submitted, one in Operating System domain and the other in Networking domain. Now both the papers can have the term ‘networks’ and ‘Linux’, but a paper containing the term networks does not mean it is a networking domain paper and a paper containing the term Linux does not indicate that it belongs to the operating system domain.

It might happen that the term network is used in the operating system paper in the context of setting up a network and in a networking domain paper also the term network can be used in the context of setting up a network. In word-content matrix method, the meaning of a term is understood by its association with the corresponding terms and as the usage of the term networking in both the operating system and network domain paper is in the context of setting up of a network, then both the papers can get categorized into network domain, which is not at all a desirable result. This kind of confusion might arise if word-content matrix is used for automatic research paper categorization for organizing a conference.

The term document matrix category is another major vector space model of semantics category. In this category the similarity between the documents is considered. In this method the frequency of all the individual terms are calculated and the terms having frequency above a threshold limit is then matched with pre-built data set or corpus. So, here if the problem that occurred with the word content matrix is considered, then even though in both the papers the term networking is used in the same meaning, that fact will not be considered. The fact that will be taken into account here is the frequency of the individual terms and it is obvious that in the network domain paper the frequency of the term network will be more and in the operating system domain paper, other terms that are more related to operating system will have a higher frequency. So there will arise no problem in categorization of the uncategorized document.

When an uncategorized research paper will be submitted for categorization then the following steps will be followed:

Step 1: Build corpora according to the need.

For research paper based on subject domain we can build corpora based on each subjects like networking, data mining, digital image processing etc.

Step 2: Read the submitted uncategorized research paper that needs to be categorized.

Step 3: Remove the stop words.

Step 4: Perform stemming on the required words of the document.

Step 5: Count the frequency of each of the terms in the document.

Step 6: Take a note of the terms having the highest term frequencies in a decreasing order.

Step 7: Omit the terms that do not cross the threshold limit.

Step 8: Match this term with the pre built corpora.

Step9: Calculate the sum of the matches of each category.

Step 10: Calculate the frequency by diving the sum of matches found by the total number of words present in that category.

Step11: Categorize the document according to the result found.

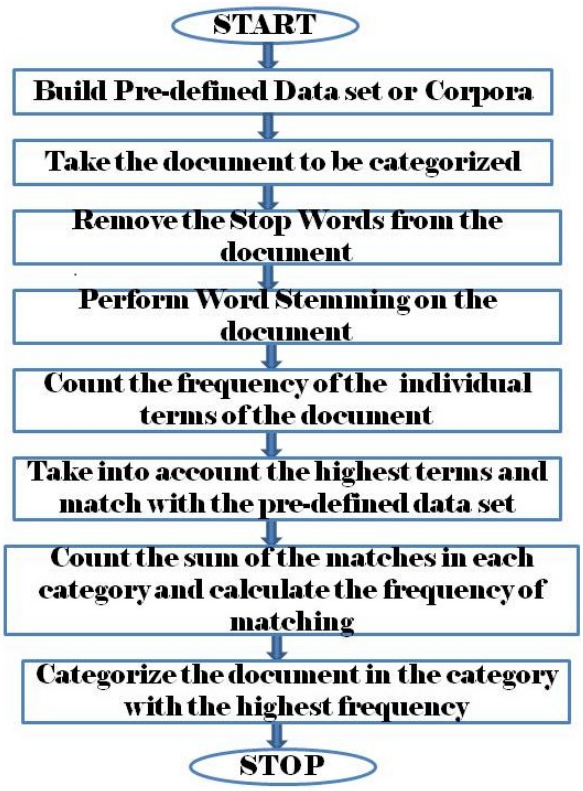

Figure 1: Flow diagram showing the steps of the proposed idea

Figure 1 is showing the proposed architecture. Using this proposed method the uncategorized research papers can be easily categorized in the proper category based on hypothesizes of the term-document matrix category of the Vector Space Models on semantics.

IV. SIMULATION RESULTS

The simulation of the above proposed idea was carried out in C# language using Microsoft Visual Studio 2008 software to verify the effectiveness of the proposed idea. Pre-defined data sets on Operating System, Networking, Wireless Sensor Networks, Software Engineering, Software Testing and Digital Image Processing were built and the submitted document was checked against those data sets. The documents to be submitted can be in .txt or .doc or .docx format. The C# code of the above proposed idea was written and uncategorized documents from different categories are submitted for categorization and the result is shown below:

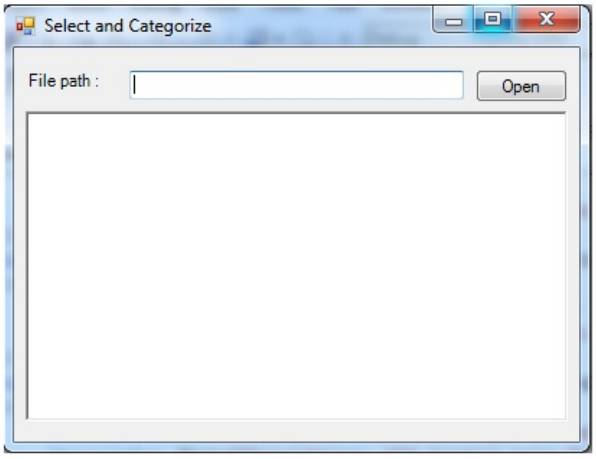

Figure 2: Asking for submission of document to be categorized.

Figure 3: Submission of document for categorization

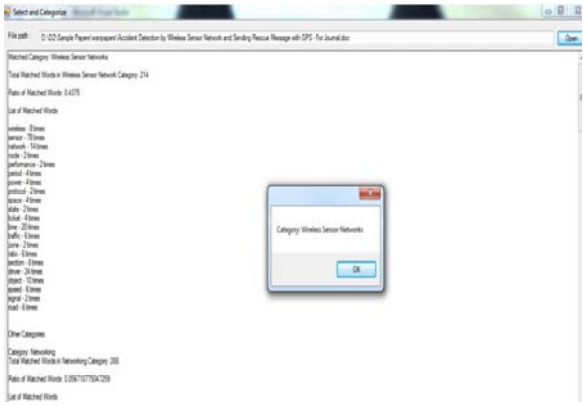

Figure 4: Showing category as Wireless Sensor Network

Figure 5: Result of Categorization

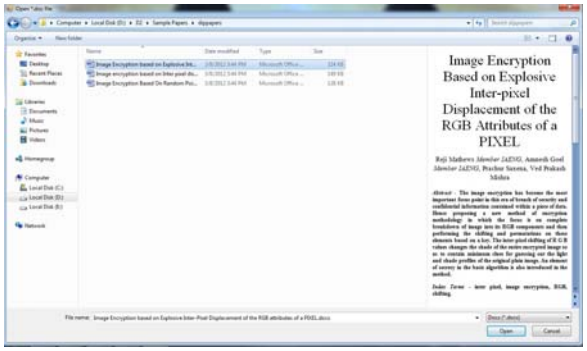

Figure 6: Submitting a different document for categorization

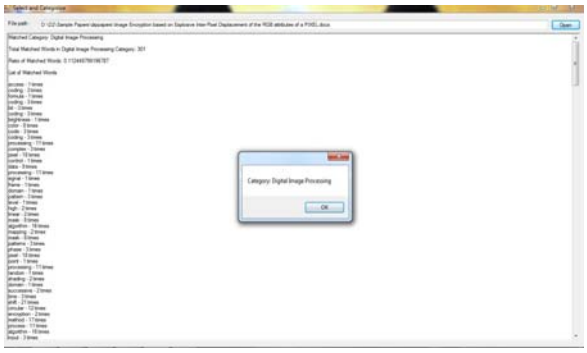

Figure 7: Showing the result and Category as Digital Image Processing

Figure 8: Result of categorization

V. CONCLUSION

From the various literature studies of the various applications of each category of the Vector Space Model of semantics, it can be said that the automatic categorization of research papers for organizing a conference can be done using any one of the three major categories of the VSM. But the term-document matrix method is comparatively more efficient than the word content matrix and the pair pattern matrix and the proposed method works appropriately based on the hypothesis of the term-document matrix category. The simulation results obtained from the proposed methodology shows successful categorization of documents into their respective categories. The range of categorization can be increased by increasing the pre- built corpora. The C# code supports documents to be submitted in .txt, .doc, .docx format but it can be optimized in the future so that it can accept documents in all other formats.

REFERENCES

[1] http://nlp.cs.nyu.edu/sk-symposium/slides/PeterTurney.pdf

[2] Manning, C. D., Raghavan, P., & Sch¨utze, H. (2008). ”Introduction to Information Retrieval.” Cambridge University Press, Cambridge, UK.

[3] Pantel, P., & Lin, D. (2002a). “Discovering word senses from text.” In the proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 613–619, Edmonton, Canada.

[4] Choi, F. Y. Y. (2000).” Advances in domain independent linear text segmentation.” In the proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, pp. 26–33.

[5] Dang, H. T., Lin, J., & Kelly, D. (2006). “Overview of the TREC 2006 question answering track.” In the proceedings of the Fifteenth Text Retrieval Conference (TREC 2006).

[6] Chu-carroll, J., & Carpenter, B. (1999). “Vector-based natural language call routing.” Computational Linguistics, 25 (3), 361–388.

[7] Rapp, R. (2003). “Word sense discovery based on sense descriptor dissimilarity.” In the proceedings of the Ninth Machine Translation Summit, pp.315–322.

[8] Sch¨utze, H. (1998). “Automatic word sense discrimination.” Computational Linguistics, 24 (1), 97–124.

[9] Curran, J. R., & Moens, M. (2002). “Improvements in automatic thesaurus extraction” .In Unsupervised Lexical Acquisition: Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX), pp. 59–66, Philadelphia, PA.

[10] Jones, M. P., & Martin, J. H. (1997). “Contextual spelling correction using latent semantic analysis.” In the proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 166–173, Washington, DC.

[11] Pennacchiotti, M., Cao, D. D., Basili, R., Croce, D., & Roth, M. (2008). “Automatic induction of FrameNet lexical units.” In the proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-08), pp. 457–465, Honolulu, Hawaii.

[12] Cao, H., Jiang, D., Pei, J., He, Q., Liao, Z., Chen, E., & Li, H. (2008). “Context-aware query suggestion by mining click-through and session data.” In the proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’08), pp. 875–883. ACM.

[13] Turney, P. D. (2006). “Similarity of semantic relations.” Computational Linguistics, 32 (3), 379–416.

[14] Lin, D., & Pantel, P. (2001). “DIRT – discovery of inference rules from text.” In the proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001, pp. 323–328.

[15] Davidov, D., & Rappoport, A. (2008). “Unsupervised discovery of generic relationships using pattern clusters and its evaluation by automatically generated SAT analogy questions.” In the proceedings of the 46th Annual Meeting of the ACL and HLT (ACL-HLT-08), pp.692–700, Columbus, Ohio.

[16] Turney, P. D. (2008b). “A uniform approach to analogies, synonyms, antonyms, and associations.” In the proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 905–912, Manchester, UK.

[17] Turney, P. D. (2008a). “The latent relation mapping engine: Algorithm and experiments.” Journal of Artificial Intelligence Research, 33, 615 -655.

Sukanya Ray is currently pursuing M.Tech in Computer Science and Engineering at Amity School of Engineering & Technology, Amity University, Noida. She has received B.Tech degree from IMPS College of Engineering and Technology, WBUT in 2010. Her current research includes in MANET, sensor network and NLP.

Ms. Nidhi Chandra has more than 7 years’ experience in academic and Software Development. She is M.Tech from CDAC NOIDA, affiliated from GGSIPU, Delhi. Presently she is working as Assistant Professor at Amity University Noida. She has Worked with Tata Unisys and CDAC Noida. Her research interest includes Natural Language Processing, Assistive Technology and Semantic Web Based Application.

