An Intelligent hybrid tool for finding and organizing relevant text

M.K.Kowar
Professor, Department of Electronics & Telecommunication
kowar_bit@Rediffmail.com

Sanjay Sharma
Reader, Department of Mathematics
ssharma_bit@rediffmail.com

Arpana Rawal
Reader, Department of Computer Science & Engineering
arpana_rawal@rediffmail.com

Ani Thomas
Reader, Department of Computer Applications
tpthomas22@yahoo.com

ABSTRACT

Recent work in intelligent text retrieval systems have shown improvements in knowledge representation techniques, irrespective of user-specific tasks viz. classification or categorization, summarization, indexing and ranking. Text miners have begun extracting and aggregating key concepts by slowly shifting from utilizing the explicitly available ontologies towards machine generated ones. In the present communication, the authors present text mining experiments in the closed-world domain to rank the documents, here chosen as academic realm. The proposal offers a two-stage hybrid tool, where a confusion matrix obtained from suitably chosen naïve-bayes classifier is used to arrive at similarity matrix, that is put to hierarchical agglomerative clustering procedures. This is extended to render an accurately precise hierarchical topic and sub-topic sequencing in the considered domain of context. The resulting accuracy of term-to-term arrangements in topic hierarchy were found promising as the same was found to be preferred, when consulted with subject experts.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering, Retrieval model and Clustering.

General Terms

Measurement, Design, Experimentation, Verification.

Keywords

Semantic concept spaces, Naïve Bayesian classifier, Confusion matrix, similarity matrix, Agglomerative clustering.

1 INTRODUCTION

In this era of information revolution, the utmost need of the hour for knowledge engineers has been to make the best of the resources. In the present context considered, academicians need to organize and re-organize their updated learning material in periodic continuum in order to make the teaching-learning process for a subject domain more captivating. Text miners look upon this task as a part and parcel of grouping conceptually related topics for any domain. This offers a more comprehensive approach to integrate and correlate the content of the given learning material in the form of a precise topic sequencing hierarchy. In this direction, an appreciable noteworthy contribution is given by the authors as described in the following section.

2 THE STATE-OF-ART

Over the years, text miners have been querying Interfaces for accessing the meaningful courseware material, to suffice the E-learning process. Several methods to machine-aided query evaluation have convinced the experts that a finitely defined concept storage space comprising of key terms lays a sound foundation to extract the underlying relevant concepts of the domain. Moreover, the previous work by the authors themselves, has shown the process of self-evolving text-learning material in the shape of precise document beds, comprising of n-gram pools and depicting content semantics as term-to-term associations [3]. An Innovative breakthrough by the authors of generating trees of concepts from self-acquired ontologies, has participated like a driving lexicon for retrieving the more-or-less precise targeted content [4]. Further, various approaches tat have attempted to rank the retrieved relevant text documents, have been observed to explore their semantics, as well [2] [7]. The semantically filtered content has also been visualized as a prelude to text document categorization and clustering tasks, among which, utilizing confusion matrix plot exploits the notion of similarities and dissimilarities among documents quite well [2].

3 SEMANTIC CONCEPT SPACES: THE SEARCH DOCUMENTS

The obvious storage structures for representation of key concepts were chosen to be lists growing in multi-dimensionality. For the example cited here, the key concepts are extracted from a particular syllabus content of Artificial Neural Networks. Initially, the syllabus strings are matched against the implicitly available ontology: the front table of contents and the back book of index of the book *"Neural Networks : Algorithms, Applications and Programming techniques"*authored by *James Freeman and David M. Strapetus*. At this juncture, the page overlap from the two offer a ready document bed as a baseline as shown in *table1, column 4*. As the precise formulation of topic and sub topic sequencing demands the degree of relevance measures among them, the authors are motivated to inherit the similar steps of content filtering by revealing it's depth of semantics, with the generation of dependency relations. The related work shows that one can feasibly weigh out the levels of useful content lying either in the vicinity of few paragraphs, few pages or the entire section by comparing the statistical measures of term occurrence and term-term co-occurrences [5]. The semantically filtered pages of the relevant text can then be categorized into section level (C_1), page level (C_2) and paragraph level (C_3) category levels of relevance as shown in *table 1, col 6*. These observed page ranges when assigned distinct vicinities give rise to separate search documents denoted from d_1 to d_7 as tabulated in *table 2*.

Table 1 : Observed .Vs. Predicted Document Relevance Using Bayesian Probabilities

Tuple-id	Syllabus term-id	Syllabus Strings	Relevant target pages	Semantically filtered page ranges	Category level of relevance	Observed Relevant documents	Predicted Relevant documents (with Bayesian classifier)
u_1	t_1	Elementary neurophysiology	8	8-17	C_1	d_1	d_5
u_2	t_1	Elementary neurophysiology	293	291-293	C_3	d_2	d_3,d_4
u_3	t_3	Processing Element	4	4-7	C_2	d_3	d_4
u_4	t_3	Processing Element	17-18	17-30	C_2	d_4	d_4
u_5	t_8	Neocognitron	373-393	373-393	C_1	d_5	d_5
u_6	t_9	Neocognitron architectecture	376	376-393	C_1	d_5	d_5
u_7	t_{10}	Neocognitron data processing	381	381-393	C_1	d_5	d_5
u_8	t_{11}	Neocognitron character recognition	5	5-7	C_3	d_3	d_3,d_4
u_9	t_{12}	Neocognitron handwritten digital recognition	7	6-7	C_3	d_3	d_3,d_4
u_{10}	t_{13}	Neural phonetic typewriter	274	274-275	C_2	d_6	d_4
u_{11}	t_{13}	Neural phonetic typewriter	283	283-286	C_3	d_7	d_3,d_4
u_{12}	t_{14}, t_{15}	Neural Network survey, Neural Network Models	3	3-7	C_3	d_3	d_3,d_4
u_{13}	t_{14}, t_{15}	Neural Network survey, Neural Network Models	41	NULL	--	---	----
u_{14}	t_{16}, t_{17}	Single layered perceptron, Multi layered perceptron	17	17-30	C_3	d_4	d_3,d_4
u_{15}	t_{16}, t_{17}	Single layered perceptron, Multi layered perceptron	21	21-30	C_2	d_4	d_4
u_{16}	t_{16}, t_{17}	Single layered perceptron, Multi layered perceptron	28	28-30	C_3	d_4	d_3,d_4
u_{17}	t_{16}, t_{17}	Single layered perceptron, Multi layered perceptron	24	24-30	C_2	d_4	d_4
u_{18}	t_{18}	XOR problem.	25 -27	25-30	C_3	d_4	d_3,d_4

Table 2. Selected Document Beds For Relevance Measures

Document bed	Section	Page Range
d_1	1.1	8-17
d_2	8.0	291-293
d_3	1.0	1-7
d_4	1.2	17-30
d_5	chapter 10	373-393
d_6	7.2.1	274-275
d_7	7.3.2	281-286

4 THE OBSERVED VS. PREDICTED DOCUMENT BEDS

After the content extraction process through semantic filtering technique, the carefully cleansed and trained documents are tested upon with a naïve bayes classifier, which are presumed to be well known to perform fairly well towards solving multi-class classification problems [6]. It first trains the model by calculating a generative document distribution P(d|u) for the observed relevance of each syllabus tuple ‘u_i’ in the document ‘d_j’ and then tests into which document does the term ‘t’ and term tuple-id ’u’ finds the predicted relevance. The Bayesian conditional probabilities for classifying of syllabus terms into one of the assigned document classes is expressed as:

$$P\left(d_k / t_i\right) = \frac{P\left(t_i / d_k\right) * P\left(d_k\right)}{\sum_{j=1}^{m} P\left(t_i / d_j\right) * P\left(d_j\right)}$$

where d_k represents k^{th} document bed and t_i the i^{th} term for which the predicted document bed is to be calculated among ‘m’ number of documents i.e : holding maximum value of $P(d_k|t_i)$. Thus, the predicted document beds for each of the tuple-id s are calculated as shown in *table 1, col 7 and 8*. The predicted document beds when compared with observed document beds are found to provide some degree of classification accuracies and classification errors. The confusion matrix being the most suitable representation for assessing the above parameters can be formulated as n x n matrix, for n no. of document samples. The diagonal elements of the confusion matrix specify the test documents correctly predicted to their true class and the non-diagonal elements specify the degree of confusion lying between any document to document pair, thus getting misclassified into each other.

5 FROM CONFUSION MATRIX TO SIMILARITY MATRIX

In the present domain of academic realm, it is the 7x7 document confusion matrix that depicts the true classifications and misclassifications among observed relevant documents when compared with predicted relevant ones as shown in table 3.

Table 3 : Confusion matrix stating mis-classifications for assigned and predicted relevant documents

Predicted doc bed Observed doc.bed	d_1	d_2	d_3	d_4	d_5	d_6	d_7
d_1	cf_{11} =0	cf_{12} =0	0	0	1	0	0
d_2	0	0	1	1	0	0	0
d_3	0	0	3	4	0	0	0
d_4	0	0	3	6	0	0	0
d_5	0	0	0	0	3	0	0
d_6	0	0	0	1	0	0	0
d_7	0	0	1	1	0	0	cf_{77}= 0

Thus, the matrix gives an insight to protrude similarities / dissimilarities of the documents in confusion space of 7 x 7 dimensions. With the convincing thought that two classes tend to go in a state of content overlap, if their test samples confuse or misclassify among each other; it may not be unreasonable to accept that the degrees of confusion is a measurable parameter for any two considered row vectors / column vectors of the confusion matrix. In other words, the ith document of the vector confuses at a degree of confusion denoted by the value, cfij with the jth document. Further, the authors have chosen a simple distance measure defined as sum of the absolute differences in coordinate values between two vectors. This accounts for formulation of pair-wise similarity measures between all test document pairs, giving the appearance of a similarity matrix as shown in figure 1.

$$\begin{array}{c|ccccccc|}
 & d1 & d2 & d3 & d4 & d5 & d6 & d7 \\
d1 & 0 & 3 & 8 & 10 & 2 & 2 & 3 \\
d2 & 3 & 0 & 5 & 7 & 5 & 1 & 0 \\
d3 & 8 & 5 & 0 & 2 & 10 & 6 & 5 \\
d4 & 10 & 7 & 2 & 0 & 12 & 8 & 7 \\
d5 & 2 & 5 & 10 & 12 & 0 & 4 & 5 \\
d6 & 2 & 1 & 6 & 8 & 4 & 0 & 1 \\
d7 & 3 & 0 & 5 & 7 & 5 & 1 & 0
\end{array}$$

Fig 1: Similarity Distance Matrix

One can always normalize the row vectors to unity for the sake of bringing about simplicity in tedious calculations. The same results in a normalized similarity matrix as shown in fig 2.

$$\begin{array}{c|ccccccc|}
 & d1 & d2 & d3 & d4 & d5 & d6 & d7 \\
d1 & 0 & 0.11 & 0.3 & 0.36 & 0.07 & 0.07 & 0.11 \\
d2 & 0.14 & 0 & 0.24 & 0.33 & 0.24 & 0.03 & 0 \\
d3 & 0.2 & 0.14 & 0 & 0.05 & 0.28 & 0.21 & 0.18 \\
d4 & 0.22 & 0.15 & 0.04 & 0 & 0.26 & 0.17 & 0.15 \\
d5 & 0.05 & 0.13 & 0.28 & 0.32 & 0 & 0.11 & 0.13 \\
d6 & 0.09 & 0.04 & 0.27 & 0.36 & 0.18 & 0 & 0.04 \\
d7 & 0.14 & 0 & 0.24 & 0.33 & 0.24 & 0.05 & 0
\end{array}$$

Fig 2: Normalized Similarity Matrix

6 CLUSTERING RESULTS AND DISCUSSIONS

The above similarity matrix composition is proceeded to be given as an input to a hierarchical Agglomerative clustering tool. The algorithm includes Amalgamation step that is carried out in varied linkage types [1]. In the proposed context, the distance measure between two clusters (initially documents themselves act independent clusters) is calculated as the average distance between all the coordinate pairs in any two considered different document clusters.

Thus, the clustered relevant documents obtained iteratively, depict the closeness of the content to be taught, thereby obtaining the correct sequence of topic learning as drafted in the dendrogram, shown in figure 3. The same was verified from the prescribed material, that had already been fed into the simulation tool as the syllabus snapshot. This reveals that, for grasping up the specific neuron model named '*Neocognitron*' and '*its design / functional details'* from the text d_5, one should go through the introductory aspects of *'a general neuron model'*, that can be extracted from the document d_1. Further, a survey of neural network models can only be started having already understood, its elementary physiology, and so lies the documents d6 following documents d_2 and d_7. Further, the functioning of a processing element as stated in d_6 should be dealt prior to the understanding of the different models of perceptrons, as sequenced in text documents d_3 and d_4.

7 ACKNOWLEDGMENTS

The authors wish their heart felt gratitude to the Management, Bhilai Institute of Technology, Durg, India for their inspiring encouragement and support towards the completion of the work.

8 REFERENCES

[1] Dunham Margret H. 2005. Data Mining: Introductory and advanced topics. Pearson Education. LPE.131-137.

[2] Godbole Shantanu. 2002. Exploiting Confusion matrices for automatic generation of topic hierarchies and scaling up multi-way classifiers. Annual Progress Report. Indian Institute of Technology . Bombay. India.

[3] Kowar M.K., Rawal Arpana, Thomas Ani and Sharma Sanjay. 2008. Fuzzy Decision Making for Automatic Answer evaluation in restricted domains. Journal Reflections' des ERA. Modinagar. India. in press.

[4] Kowar M.K., Rawal Arpana, Thomas Ani and Sharma Sanjay. 2007. An Automated Tool for Relevancy Ranking of Text Documents using Bootstrapping Semantics. In Proceedings of National Conference on Technological Revolution in Application Development & Intelligent Systems, (Shri Shankaracharya College of Engineering & Technology, Bhilai, Chhattisgarh, India, October 26-27 2007). TECHNOVISION' 07. 41-47.

[5] Kowar M.K., Rawal Arpana, Thomas Ani and Sharma Sanjay. 2007. Learning Ontologies and Semantic Concept Spaces for Automatic Document Relevance Ranking. In Proceedings of 2nd International Conference on Resource Utilization and Intelligent Systems, (Kongu Engineering College, Perundurai, Erode, TN. India, January 3-5 2008). INCRUIS'08. 591-595.

[6] Timothy J. Ross .1997. Fuzzy Logic with Engineering Applications. Mc. Graw Hill, Inc. International Edition. 102-105, 317-322.

[7] Stumme Gred, Tane Julien, Schmitz Christoph, Staab Steffen and Studer Rudi .2003. The Courseware Watchdog – an Ontology-based tool for finding and organizing learning material. Learning Lab Lower Saxony (L3S), Hannover, Germany; www.learninglab.de and Institute for Applied Informatics and Formal Description Methods (AIFB), University of Karlsruhe, Karlsruhe, Germany; www.aifb.uni-karlsruhe.de/WBS .