TEXT PROCESSING IN INFORMATION RETRIEVAL SYSTEM USING VECTOR SPACE MODEL

1. Premalatha.R
Department of Computer Science and Engineering,
Sri Sairam Institute of Technology,
Chennai
rprema79@rediffmail.com

2. Srinivasan.S
Head, Department of Computer Science and Engineering,
Anna University,
Madurai
sriniss@yahoo.com

Abstract— **Intelligent information retrieval is an important area in computer application in the 21st century. In Tamil documents, Morphology (separating noun and verb) concept is used to retrieve the text. In this paper we use new approach on text processing in the information retrieval system. So we can widen our search criteria namely, Vowel – Kuril (Short), Nedil (Long); Consonant - Vallinam (Hard), Mellinam (Soft) and Idaiyinam (Medium). So it would not wait for the entire word to enter; perhaps the searching process starts immediately after the first letter is entered, because the Database table is segregated into 5 components rather than a single Database table. So to minimise the time constraint, memory space and to do a smart search a new IR system is introduced. In the proposed system, searches can be divided into three categorise, namely (i) Main topic search (ii) Subtitle search and (iii) Keyword search. So the system would search quickly and retrieve required information only. In addition to that, every poem is displayed with related pictures. So users will show more interest and desire to read those poems. In the classical system, the user should give the exact word to retrieve the information. But in the proposed system the misspelled word could be corrected and the information can be retrieved, because internally the system has its own spell checker. This would be useful for Tamil literates, Tamil students, Tamil scholars, etc.**

Keywords—Text processing, Information Retrieval, Vector Space Model, Tamil Language.

I. INTRODUCTION

For thousands of years people have realized the importance of archiving and finding information. With the advent of computers, it became possible to store large amount of information and finding useful information from such collections which became a necessity. The field of IR was born in the 1950s out of this necessity Several IR Systems are used on an everyday basis by a wide variety of users. Need of information retrieval arises in Tamil literary documents from the ancient era to the latest, which helps in sharing the data through the internet. This paper proposes a new approach for text processing in information retrieval of Tamil literary document.

In the Existing system, the system [4] read the whole word and it identifies whether the word is a noun or a verb. Finally it finds the word from the database. In the past it took a lot of time for retrieval. Because, it performs the searching process after the entire word is entered. But in the proposed system, the system would not wait for the entire word to enter; perhaps the searching process starts immediately after the first letter is entered. Because the system performs the search processes for every letter of a word. This approach considerably increases the speed of the retrieval process.

Tamil language:

Tamil has a large and extremely rich body of ancient literature. It is one of the primary independent sources of modern Indian culture and tradition. Tamil is the official language of the Indian state of Tamil Nadu and one of the 22 languages under schedule 8 of the constitution of India. It is also one of the official languages of the Sri Lanka and Singapore and union territories of Pondicherry and the Andaman & Nicobar Islands.

II. RELATED WORK

Language occupies an important position in the history of Indian cultural traditions Tamil became the first legally recognized Classical language of India. Information retrieval [2] of Tamil literature is a difficult work to do because it was used in olden period Tamil format and it was on poetry format as well. Generally morphology approaches [3] are used for information retrieval of Tamil documents from the likes of Rajendran, 2001 Anand kumar M, 2009, etc., . An IR system returns a list of long documents to a user query. The construction and use of exploration models and search indices consumes processing time, memory, and disk space. Furthermore, in real systems any search and exploration methods must be computationally efficient. In particular, the delay perceived by the users is critical. It is therefore important to develop methods that can speed up the search process while maintaining high perceived quality, particularly in the range of high precision and low recall which is most crucial in actual user settings. The proposed system is a new approach of text processing in information retrieval using vector space model that were organized for exploration of Tamil document collections.

ISBN No.978-1-4799-3834-6/14/$31.00©2014 IEEE

III. TAMIL SCRIPTS

It is inherited from brahmi script. Native grammarians classify [1] Tamil phonemes into vowels, consonants, and a secondary character (āytam). It has 12 vowels and 18 consonants. These combined with each other to yield 216 characters and 1 special character (Aayitha Ezhuthu) counting to a total of (12+18+216+1) 247 characters.

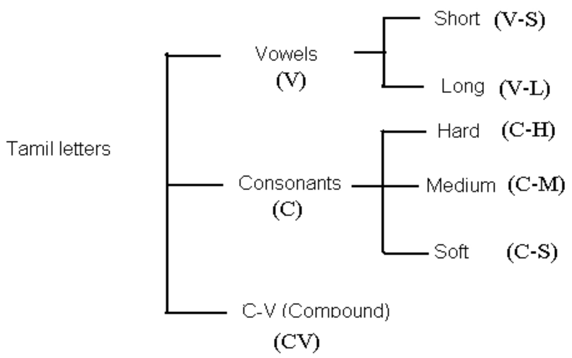


Fig. 1 Classification of Tamil Letters

A. *Vowels*

Vowels are also called as uyirezhuttu (uyir – life, ezhuttu – letter) in Tamil. These vowels are classified into Short (kuril) and Long (Nedil), five of each type and two diphthongs, /ai/ and /au/. The long vowels are about twice as long as the short vowels. The diphthongs are usually pronounced about 1.5 times as long as the short vowels.

B. *Consonants*

Consonants are known as Meyyezhuttu (mey—body, ezhuttu—letters) in Tamil. It is classified into three categories Hard (Vallinam), Medium (Idaiyinam) and Soft (Mellinam) and six in each category. The classification is based on the place of articulation

TABLE I
VOWEL CONSONANT TABLE

Tamil letters	Vowels (V)	Short (V-S)	அ	இ	உ	எ	ஒ	ஐ
		Long (V-L)	ஆ	ஈ	ஊ	ஏ	ஓ	ஔ
	Consonants (C)	Hard (C-H)	க்	ச்	ட்	த்	ப்	ற்
		Medium (C-M)	ய்	ர்	ள்	வ்	ழ்	ல்
		Soft (C-S)	ங்	ஞ்	ண்	ந்	ம்	ன்

IV. SYSTEM ARCHITECTURE

In this system, 3 phases are used. They are (i) Classification and Indexing (ii) Text Processor and (iii) Retrieval Unit.The system architecture as shown in figure 2.

A. *Classification and Indexing:*

Most of the ancient Tamil literatures are rendered in the form of poetries. The critical edition of ancient Tamil works include 41 works namely 1) Thirukkural 2) Pura naanooru 3) Aga naanooru 4) Silapathigaram 5) Seevaga chinathamani 6) Manimegalai 7) Kundalakesi 8) Valayapathi 9) Padhinen Mel kanakku (18 Upper Classics) 10) Padhinen Keezh kanakku (18 Lower Classics) etc. Since most of the Ancient Tamil works are in poetry and in anthology forms, a Main class is derived with 41 categories [10] and each category has various sub divisions.
The following tasks are performed in this module.

1. Major groups in ancient tamil literature are identified.
2. Identified groups are classified into various categories.
3. Main topics are identified as headings of the categories.
4. Topics are hierarchically arranged along with the sub titles and
5. All the key words are arranged along with the topics and sub titles.

Database table is segregated into 5 components namely Vowel–Short **(V-S)**, Vowel-Long **(V-L)**, Consonant–Hard**(C-H)**, Consonant–Medium**(C-M)**, Consonant–Soft**(C-S)**. Database Table segregation can be done by the following steps:

1. Select the Number of documents as search data.
2. Read every document term by term
3. Apply the CV test to check if the first letter of the term is a Vowel (V) or a Consonant (C). If it is vowel, then check whether it is Short or Long. After identifying the type, update the term to the corresponding database table.
4. Similarly, check whether the first letter of the term is a Consonant. Then check if it is Hard or Medium or Soft. After identifying the type update the term to the corresponding database table.

ISBN No.978-1-4799-3834-6/14/$31.00©2014 IEEE

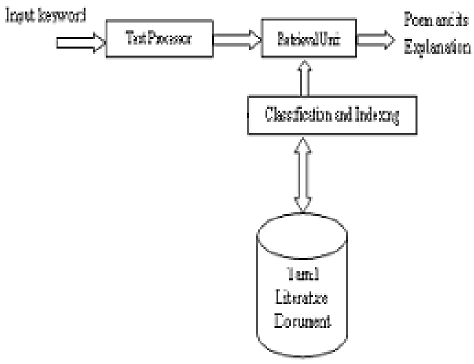


Fig. 2 System Architecture

B. *Text Processor:*

Every letter has been processed and evaluated by the Text processor based on Tamil phonemes [2]. If a key is pressed through the Tamil virtual keyboard, the processor would identify the letter through any one of the following – Vowel (V) or Consonant(C). Vowels are again classified into two types, Short(S) and Long (L). Similarly consonants are classified into three classes with 6 in each class and are called Hard (H), Medium (M) and Soft(S).

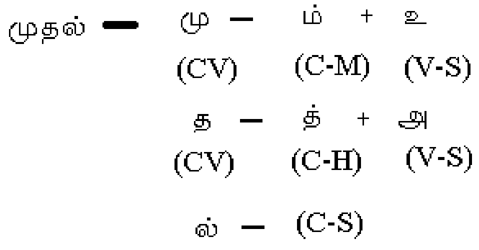


Fig. 3 Input keyword 1

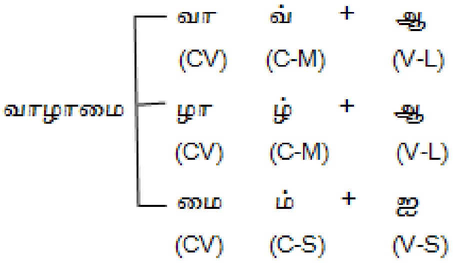
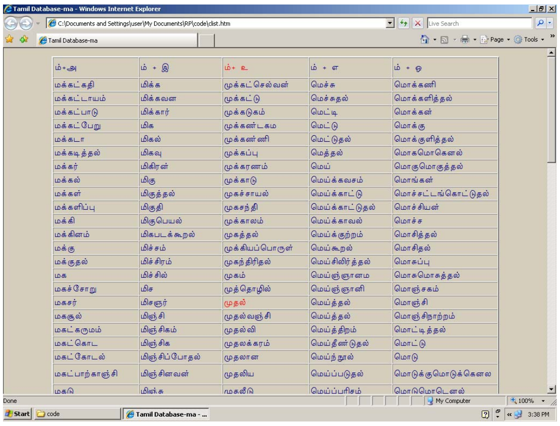


Fig. 4 Input keyword 2

Once the letter is identified, it locates the letter from the Database. The cycle would continue until all the letters in the word are processed. The index position should change for every additional letter until the complete word is entered.

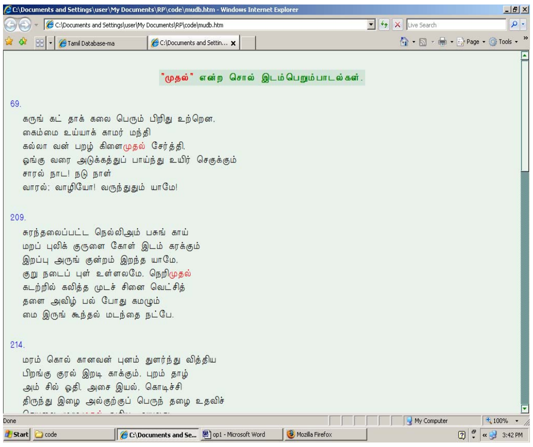

Tamil Database-ma - Windows Internet Explorer
C:\Documents and Settings\user\My Documents\RP\code\clist.htm
Live Search
Page
Tools
Tamil Database-ma
Done
My Computer
100%
Start
code
Tamil Database-ma - ...
3:38 PM

ம்+அ	ம் + இ	ம்+ உ	ம் + எ	ம் + ஒ
மக்கட்கதி	மிக்க	முக்கட்செல்வன்	மெச்சு	மொக்கணி
மக்கட்டாயம்	மிக்கவன	முக்கட்டு	மெச்சுதல்	மொக்களித்தல்
மக்கட்பாடு	மிக்கார்	முக்கடுகம்	மெட்டி	மொக்கன்
மக்கட்பேறு	மிக	முக்கண்டகம	மெட்டு	மொக்கு
மக்கடா	மிகல்	முக்கண்ணி	மெட்டுதல்	மொக்குளித்தல்
மக்கடித்தல்	மிகவு	முக்கப்பு	மெத்தல்	மொகமொகெனல்
மக்கர்	மிகிரன்	முக்கரணம்	மெய்	மொகுமொகுத்தல்
மக்கல்	மிகு	முக்காடு	மெய்க்கவசம்	மொங்கன்
மக்கள்	மிகுத்தல்	முகச்சாயல்	மெய்க்காட்டு	மொச்சட்டங்கொட்டுதல்
மக்களிப்பு	மிகுதி	முகசந்தி	மெய்க்காட்டுதல்	மொச்சியன்
மக்கி	மிகுபெயல்	முக்காலம்	மெய்க்காவல்	மொச்ச
மக்கினம்	மிகபடக்கூறல்	முகத்தல்	மெய்க்குற்றம்	மொசித்தல்
மக்கு	மிச்சம்	முக்கியப்பொருள்	மெய்கூறல்	மொசிதல்
மக்குதல்	மிச்சிரம்	முகந்திரிதல்	மெய்சிலிர்த்தல்	மொசுப்பு
மக	மிச்சில்	முகம்	மெய்ஞ்ஞானம	மொசுமொசுத்தல்
மகச்சோறு	மிச	முத்தொழில்	மெய்ஞ்ஞானி	மொஞ்சகம்
மகசர்	மிசஞர்	முதல்	மெய்த்தல்	மொஞ்சி
மகசூல்	மிஞ்சி	முதல்வஞ்சி	மெய்த்தல்	மொஞ்சிநாற்றம்
மகட்கருமம்	மிஞ்சிகம்	முதல்வி	மெய்த்திறம்	மொட்டித்தல்
மகட்கொட	மிஞ்சிக	முதலக்கரம்	மெய்தீண்டுதல்	மொட்டு
மகட்கோடல்	மிஞ்சிப்போதல்	முதலான	மெய்ந்நூல்	மொடு
மகட்பாற்காஞ்சி	மிஞ்சினவன்	முதலிய	மெய்ப்படுதல்	மொடுக்குமொடுக்கெனல
மகடு	மிஞ்சு	முகவீடு	மெய்ப்பரிசம்	மொடுமொடெனல்

Fig. 5 ம் Database Table

C. *Retrieval Unit:*

In the retrieval module [12], once the word is found in indices, it retrieves the poem and its explanations from the documents using vector space model [3].


C:\Documents and Settings\user\My Documents\RP\code\mudb.htm - Windows Internet Explorer
C:\Documents and Settings\user\My Documents\RP\code\mudb.htm
Live Search
Tamil Database-ma
C:\Documents and Settin...
Page
Tools
Done
My Computer
100%
Start
code
C:\Documents and Se...
op1 - Microsoft Word
Mozilla Firefox
3:42 PM

"முதல்" என்ற சொல் இடம்பெறும்பாடல்கள்

69.

கருங் கட் தாக் கலை பெரும் பிறிது உற்றென,
கைம்மை உய்யாக் காமர் மந்தி
கல்லா வன் பறழ் கிளைமுதல் சேர்த்தி,
ஓங்கு வரை அடுக்கத்துப் பாய்ந்து உயிர் செகுக்கும்
சாரல் நாட! நடு நாள்
வாரல்; வாழியோ! வருந்துதும் யாமே!

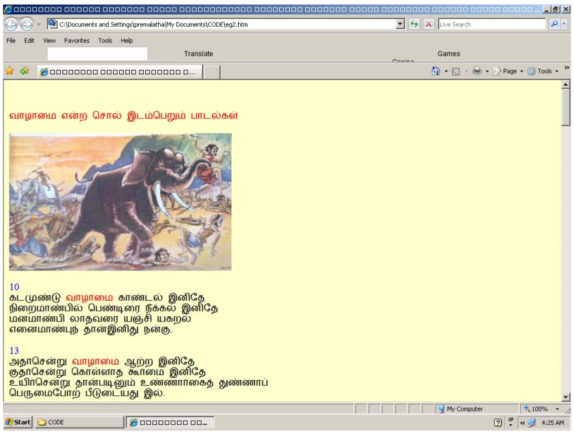
209.

சுரந்தலைப்பட்ட நெல்லிஅம் பசுங் காய்
மறப் புலிக் குருளை கோள் இடம் கரக்கும்
இறப்பு அருங் குன்றம் இறந்த யாமே,
குறு நடைப் புள் உள்ளலமே, நெறிமுதல்
கடற்றில் கலித்த முடச் சினை வெட்சித்
தளை அவிழ் பல் போது கமழும்
மை இருங் கூந்தல் மடந்தை நட்பே.

214.

மரம் கொல் கானவன் புனம் துளர்ந்து வித்திய
பிறங்கு குரல் இறடி காக்கும், புறம் தாழ்
அம் சில் ஓதி, அசை இயல், கொடிச்சி
திருந்து இழை அல்குற்குப் பெருந் தழை உதவிச்

Fig. 6 Screenshot of Input keyword 1



வாழாமை என்ற சொல் இடம்பெறும் பாடல்கள்

10
கடமுண்டு வாழாமை காண்டல் இனிதே
நிறைமாண்பில் பெண்டிரை நீக்கல் இனிதே
மனமாண்பி லாதவரை யஞ்சி யகறல்
எனைமாண்புந் தான்இனிது நன்கு.

13
அதர்சென்று வாழாமை ஆற்ற இனிதே
குதர்சென்று கொள்ளாத கூர்மை இனிதே
உயிர்சென்று தான்படினும் உண்ணார்கைத் துண்ணாப்
பெருமைபோற் பீடுடையது இல்.

Fig. 7 Screenshot of Input keyword 2

The performance of the search can be increased by splitting up the category by 3 ways namely (I) Main topics searches (2) Subtitles searches and (3) Keyword searches. For example, if the input is given from “Main topics”, then the subtitle and keyword would be skipped. Examples are shown in Figure 8 and 9.

Main Category : சீவகசிந்தாமணி
Main Topic : நாமகள் இலம்பகம்
Subtitle : நாட்டு வளம்
Keyword : திசை

Fig. 8 Example 1

Main Category : திருக்குறள்
Main Topic : பாயிர இயல்
Subtitle : கடவுள் வாழ்த்து
Keyword : இறைவன்

Fig. 9 Example 2

Fig. 10 Sample pictures for Main Topics

D. *Spell Checker:*

The spell checker detects the mistakes and prompts the user with a set of suggestions, which will aid the correction of the misspelled word. Generally the user would not enter the first letter wrongly, but he might do in the middle of the word for the letters like

1) ர ற 2) ல ள ழ 3) ன ண ந

So the spell checker is designed in such a way that the first letter would be skipped and the rest of the letters in the word should undergo the spell-check. Ideally it skips the first letter of the word and does a spell check for the rest of the letters of the word. There are two cases involved in spell checker.

Case 1: If there is single combination available for the input in the DB, it gets the word from the DB and replaces it.

Case 2: If there is more than one combination, it would list all possible combinations to select. Look at these examples below:

Case: 1 கள்வி → கல்வி புகல், புகள் → புகழ்
அரிதல் → அறிதல்

Case: 2 பளம் → பலம் , பழம்

Fig. 8 Examples for Case 1 and Case 2

V. VECTOR SPACE MODEL

There are three classic models in the information retrieval given by: (1) Boolean model in which the document and queryare represented as sets of index terms; this model is a set theoretic. (2) Probabilistic model in which the framed work for modelling document and query representation is based on probability theory; this model is probabilistic (3) Vector space model in which the documents and query are represented as vectors in t-dimensional space. Thus, this model is algebraic and it is of main concern of our study.

In the vector space model, we represent documents as vectors. The success of the vector space method is based on term weighting. There has been much research on term weighting techniques but little consensus on which method is best.

A. *Term Weighting:*

Term weighting is an important aspect of modern text retrieval systems. Terms are words, phrases, or any other indexing units used to identify the contents of a text. Since different terms have different importance in a text, an important indicator – the Term Weight – is associated with every term.

$$\begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ \vdots \\ D_n \end{array} \left(\begin{array}{cccc} T_1 & T_2 & \cdots & T_t \\ w_{11} & w_{21} & \cdots & w_{t1} \\ w_{12} & w_{22} & \cdots & w_{t2} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ w_{1n} & w_{2n} & \cdots & w_{tn} \end{array} \right)$$

1) Term Frequency: It is defined in the simplest case as the number of occurrences of the term in the document. We would like to compute a score between a query term t and a document d, based on the weight of t in d. The simplest approach is to assign the weight to be equal to the number of occurrences of term t in document d. This weighting scheme is referred to as term frequency and is denoted $\mathbf{tf}_{t,d}$, with the subscripts denoting the term and the document in order.

ISBN No.978-1-4799-3834-6/14/$31.00©2014 IEEE

2)*Document Frequency:* It is defined to be the number of documents in the collection that contain a term t and is denoted **df_t.**

3)*Inverse document frequency:* It is a measure of the general importance of the term (obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient).

$$\mathrm{idf}_t = \log \frac{N}{\mathrm{df}_t}.$$

Where, N : total number of documents in a collection
df_t: Document frequency

B. Tf-idf Weighting:

We now combine the definitions of term frequency and inverse document frequency, to produce a composite weight for each term in each document. The tf-idf weighting scheme assigns to term t a weight in document d given by

$$\textbf{tf-idf}_{t,d} = \textbf{tf}_{t,d} \times \textbf{idf}_t$$

In other words, tf-$\mathrm{idf}_{t,d}$ assigns to term t a weight in document d that is

1. highest when t occurs many times within a small number of documents (thus lending high discriminating power to those documents).

2. Lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal);

3. Lowest when the term occurs in virtually all documents.

VI. EVALUATION

The objective evaluation of search effectiveness has been a cornerstone of IR. Progress in the field critically depends upon experimenting with new ideas and evaluating the effects of these ideas, especially given the experimental nature of the field. From the early years, it was evident to researchers in the community that objective evaluation of search techniques would play a key role in the field. The two desired properties that have been accepted by the research community for measurements of search effectiveness are *recall[3]*: the proportion of relevant documents retrieved by the system; and *precision*: the proportion of retrieved documents those are relevant.

Experiment Results:

Information retrieval systems Comparison:

❖ Recall improvement:

Input String 1: + 0.14.
Input String 2: + 0.50.

❖ Precision improvement:

Input String 1: + 0.51.
Input String 2: + 0.44.

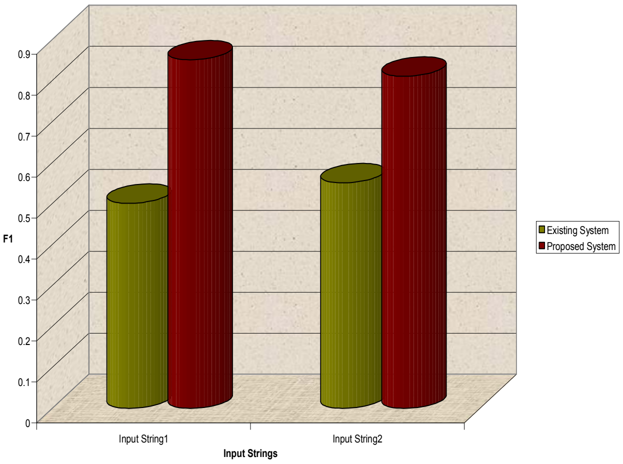


Figure 4-PERFORMANCE COMPARISON

VII. CONCLUSION

The proposed system is focuses on information retrieval for Tamil literary document. As the pictures are included along with the poems it would be easier and interactive for the learners to remember and understand easily. The proposed system would be pretty much helpful for all Tamil literates and students to search and learn.. Also performance tuning has also been done in the system. This system currently supports 41 categories only. In addition to this we can further add more documents in future. However this concept will be designed more useful to the users in future.

REFERENCES

[1] Abirami.S and D. Manjula, “Feature string-based intelligent information retrieval from Tamil document images”, International Journal of Computer Applications in Technology, Publication, Volume 35, pp. 150-164, 2009.

[2] Amit Singhal, Google, “Modern Information Retrieval: A Brief Overview, IEEE transactions”, IEEE Computer Society Technical Committee on Data Engineering, 2001.

[3] Anand kumar M, Dhanalakshmi V, Rajendran S, Soman K P, “A Novel Approach to Morphological Analysis for Tamil Language”, Internet Tamil Conference, 2009.

[4] Anandan.P, T.V. Geetha, and Ranjani Parthasarathi (2001), “Morphological Generator for Tamil”, Tamil Inaiyam 2001, Malaysia.

[5] Bayard, R. J, Ma. Y. Srikant, “Scaling up all pairs similarity search”. In Proceedings of the 16th international conference on World Wide Web (WWW '07), pp. 131-140, New York, 2007.

Chris Buckley, Gerard Salton, and James Allan. Automatic retrieval with locality information using SMART. In Proceedings of the First Text REtrieval Conference (TREC-1), pages 59–72. NIST Special Publication 500-207, March 1993.

[6] N. Guarino, C. Masolo, G Vetere, OntoSeek: Content-Based Access to the Web, IEEE Intelligent Systems, pp 70-80, May/June 1999.

[7] Holger, Billhardt, Victor Maojo, “A context vector model for information retrieval”, Journal of the American Society for Information Science and Technology, Volume 53, Pages: 236 - 249, Year of Publication: 2002.

[8] Massimo Melucci, “A basis for information retrieval in context”, ACM Transactions on Information Systems (TOIS), volume.26, p.1-41, June 2000.

[9] Pablo Castells, Miriam Ferna´ndez, and David Vallet, “An Adaptation of the Vector-Space Model for Ontology-Based Information

[10] Retrieval” IEEE Transactions on Knowledge and Data Engineering, Volume.19, No.2, February 2007, ISSN: 0975-5462 4432

[11] Rajan.K, Ramalingam.V, M.Ganesh, “Automatic classification of Tamil documents using vector space model and artificial neural network”, Expert Systems with Applications: An International Journal, Volume 36, and Issue 8 Pages: 10914-10918, 2009

[12] Ray Larson, Marc Davis. SIMS 202: “Information Organization and Retrieval”. UC Berkeley SIMS, Lecture 18: Vector Representation, 2002.

[13] Wan, V. N. Anh, I. Takigawa, and H. Mamitsuka, “Combining vector-space and word-based aspect models for passage retrieval”. In Proc. 15th Text Retrieval Conference (TREC 2006), Special Publication 500-272, November 2006.

[14] Yunjae Jung, Haesun Park and Ding-zhu Du. An effective Term-Weighting Scheme for Information Retrieval. Technical Report TR 00-008, Department of Computer Science and Engineering, University of Minnesonta, Minneapolis, USA 2000

[15] Zachary G. Ives. Information Retrieval. University of Pennsylvania, CSE 455-Internet and Web Systems, 2004