(IJAER) 2018, Vol. No. 15, Issue No. V, May **e-ISSN: 2231-5152, p-ISSN: 2454-1796**

SEARCH STRATEGYIMPROVING IN SEARCH ENGINE

***DIVYANSH GUPTA, **UJJWAL JAIN**

**N.K. Bagrodia Public School, Dwarka*

***Sachdeva Public School, Pitampura*

ABSTRACT

Users on the internet uses search engine to find information of their interest. However current search engines on webreturn answer to a query of user independent of user's requirement for the information. In this paper our aim is touse a new technique called probabilistic latent isaccurate than previously used techniques by various search engines. Our main focus in this paper is on therequirement for more accurate search results by meta search engine. In comparison wfor searching like LSA, which perform singular value decomposition of cothis paper, relies on mixture decomposition derivedfrom latent class model. Results obtained by PLSA in search for query shows that this technique gives more accurate results in searching most relevant document from a givencorpus for a query of user.

Keyword *- Meta search engine, Indexing Query, Vector Space Model, Latent Semantic Indexing, WordMatrices, Probabilistic Latent Semantic Analysis, Expectation Maximization algorithm.*

INTRODUCTION

As of late's web is developing quickly and ubiquity is expanded up to a point that each individual thinks about it and make its utilization for various purposes. A few people utilize web to think about something new in the present condition while others utilize it as a methods for excitement. utilization of web isn't restricted to diversion however it can likewise be utilized to lead examine related work, such as finding and perusing most recent explores on current patterns. Web is additionally utilized for getting most recent news. A large number of site pages are added to this web with every human need. A review by Google speak to that there are one trillion one of a kind URL's on the web. The usage of web crawler makes the way toward looking through a portion of the themes of client enthusiasm for a simple way. Queering the singe a specific theme would recover the outcomes from the web and exhibited to the web clients. Since there are expansive number of website pages on the web and hence result acquired are likewise tremendous. Client gets all that could possibly be needed web connects subsequently delivered via web crawler and squanders their valuable time in exploring through undesirable connections, looking through the required one. The principle purpose behind this is the Search Engine do the ordering of the pages based on content entered by client. With the end goal to beat this weakness, we have to execute a strategy that will enable the client to locate the important words, beginning from the few words that they may really know. At the end of the day, we have to center around the semantic of words entered by client. This exploration paper shows another methodology that depends on a few calculations which considers semantic parts of content and uses them to actualize a Meta web crawler that will give client suitable outcomes in their inquiry for applicable data. For PCs to cooperate all the more normally with people, It is important to manage clients asks for that don't have clear significance or we can state that arrangement with unfeasible client demands is vital. It is a vital need to perceive the contrast between what a client may state or do and what he/she really need and planned for. A procedure of data recovery utilizing web indexes involves following advances.

1. By nlp, for e.g. user provides some keywords to web search engine and expects that it will ret relevant data in response to their query.

2. Web Search-Engines make use of a special program called spider, travels the web from one page to another. It travels the popular sites on the internet and then follows each link available at that site. This program saves all the words and their respective position on the visited web

3. After collecting and storing all the data, search engines build an index to store that data so that a user can access pages quickly. The technique used by various internet searcher for ordering is extraordinary and in this way the outcome delivered by a various web search tool for a similar question is unique. Vital focuses considered amid ordering process include the recurrence of a term showing up in a page, segment of a site page where that term shows up, text dimension of a term. Ordering data is encoded into lessened size to accelerate the reaction time of the specific hunt motor, and afterward it is put away into the database

PROBLEM DEFINITION AND STATEMENT

Techniques adopted for by meta search engine insearching an archive significant to client inquiry not giving the palatable outcomes to the clients. The important behind the methods utilized is either extricating the inclinations given by the client or keeping up client profile. Some iterative calculations are connected on web index results to refine the outcomes properly and all the more precisely. The yield of these calculations gives the arrangement of the issue definition clarified here in this area. The primary purpose of reasoning is the decision of proper calculation for enhancing the hunting procedure of Meta web search tool [1]. When workingwith search engine users faces a common problem ofnot getting the desired information quickly in an easyway. The main problem is that when user enters sometext keyword in search engine, it will return a list ofvarious web pages on the basis of keyword typed bythe user. Usually search engine does not respond withonly the result that user actually needed, instead itgives lots of undesirable web page links and userwastes their precious time in navigating from oneweb page to another in search for the document whatthey actually want.For improving the search strategy keywords typed bythe user in search for the information what theyneeded is also an important issue. Many internetusers want information of their interest on web, butthey do not know how to get that information fast in an easy way. The choice of keyword typed by user isalso a critical issue. Another aspect of problemdefinition depends on the ability of search engine torespond with appropriate search result. Not anysearch engine discovered yet, is capable of coveringeven a half portion of the web pages available on thenet [2]. Some search engines give the web pages thatare visited many times and thus the required pagedoes not come in front of the user and they makesearch again and again, but always gets the sameresult for a given keyword through a specific searchengine. An even sometimes search engine gives suchweb page links in results which contain obsolete ordead link [3].A study was performed to evaluate the similaritiesand differences between the search results given bythe three search engines named Google, Yahoo,Ask Jeeves, and this procedure is named assessing covering among first page aftereffects of the previously mentioned web indexes. This examination uncovers that 92.53 percent of URL is recovered by one web search tool no one but (which could be any out of the three), 5.22 percent URLs are shared by two, while 2.02percent and 0.21 percent of URLs were retrieved byall three search engines. This small percentage ofoverlapping between SEs shows that there is asignificant difference in search strategy of all SEs.

PROPOSED FRAMEWORK

Our proposed model is based on the Vector SpaceModel and later we further extend it to the PLSA
(Probabilistic Latent Semantic Analysis) model andthen examine how these models worked to performquery expansion. On Internet different content recovery procedures depend on ordering of content watchword, since catchphrase alone isn't equipped for catching the entire report content fittingly, theperformance of retrieval strategy becomes poor. Butusing the indexing mechanism of keywords we canprocess large corpa of document in an efficient way.When identification of significant index word isfinished one of the two information retrieval model isused to match query to document named statisticalmodel or Boolean model. Statistical model gives thesimilarities between query and document whileBoolean model matches to an extent up to which theword satisfies Boolean expression. In 1975 Gerald Slaton [4] gives a model named "Vector Space Model" which maps the record in n-dimensional space. Where n is the quantity of various words (w1,w2, w3 … .wn) which contains the entire vocabulary of the corpus or content accumulation. Each measurement relates to a different term. In the event that a term exits in the report, its incentive in the vector is non zero. Vector activities can be utilized to contrast record and questions. In vector space display each report is considered as a vector as D1, D2, D3, D4, ,…………. Dr,
Where r is the total number of document in corpa.

Representation of document vector is
Dir = (d1r, d2r, d3r,…………………. ,
dnr)

dir represents the ith component of rth document vector.

CONCEPT OF VECTOR SPACE MODEL

Vector Space Model is an arithmetical model for speaking to content reports as vectors of identifiers. It is utilized in data sifting. Generally, this model is utilized where archives are set in term – space. Question is additionally similar to a short archive. Thismodel is required to find the most relevant documentfor the given query. In this model computation oflikenesses between gathering of reports and question is performed first and afterward restores the most precisely coordinating archives [4]. Similitudes are figured on premise of different diverse components. One of them, every now and again utilized likeness factor is the cosine comparability. Closeness between report vector and inquiry vector can be figured by, contrasting the deviation of points between each archive vector and the first question vector. By and by it is simpler to figure the cosine of edge between the vectors, rather than edges itself.cos ı= Q*D/|Q|*|D|The expression shows the cosine angles betweendocument vector D and query vector Q. If twodocuments are neighbors of each other in term space,then they would be considered relevant with eachother. By applying different similarity measures tocompare queries to terms and documents, propertiesof the record accumulation can be accentuated or deemphasized. For instance, dab item similitude measure finds the Euclidean separation between the question and a term or record in the space. Too thecosine similarity is mentioned above. Here someother factors are also mentioned for measuringsimilarity between document vector and query vector[5].

Table 1: Similarity measures of VSM

Similarity Measure	Evaluation of binary term vector
Cosine similarity	$\cos\theta = Q*D/\|Q\|*\|D\|$
Inner product	$\Sigma Qj*Dj$
Dice coefficient	$2\Sigma Qj*Dj/\{\Sigma Qj^2 + \Sigma Dj^2\}$
Jaccard coefficient	$\Sigma Qj*Dj/\{\Sigma Qj^2 + \Sigma Dj^2 - \Sigma Qj*Dj\}$

Every component of document vector is associatedwith numeric factor and that numeric factor is calledweight of the respective word or term in document.Weight associated with word wi, can be replaced byterm frequency (tfi).Here some advantages of Vector Space model overBoolean model are listed below

1. VMS is a simple model based on linearalgebra.
2. Term frequency is not binary.
3. VMS allows for calculating a continuousdegree of similarity between queries anddocuments.
4. It allows ranking of documents based ontheir possible relevance.Some limitations of VMS are mentioned below.

1. A long archive is inadequately spoken to in light of the fact that they have poor comparability esteems.
2. Inquiry catchphrases should decisively coordinate record terms.
3. Semantic affectability: records with comparable setting however extraordinary term vocabulary won't be related, bringing about a false negative match.
4. The request of term showing up in the archive has lost in vector space portrayal.
5. Weighting is instinctive however not exceptionally formal.

CONCEPT OF PROPOSED TECHNIQUE(PLSA)

Th. Hofmann presented a statistical view on LSA,which formulate the new model called ProbabilisticLatent Semantics Analysis model [6][7], whichprovide probabilistic approach for discovering latentvariables, which has a statistical foundation. Thebasic of PLSA is a latent class statistical mixturemodel named Aspect model. This aspect modelassumes that there is a set of hidden factorsunderlying the co-occurrences between twodocuments. PLSA uses Expectation-Maximization(EM) [8] to estimate the probability values thatmeasure the relationship between the hidden factorsand the two sets of documents. In this model werepresents the hidden class variable h € H = {h1, h2,h3,…………}, document d € D = { d1, d2, d3,…………} andwords w € W = {w1, w2, w3,…………}.Some parameters of this model can bedefined in the following way [9]:

P (*d*) = Probability of selecting a document*d*,
P (h |*d*) = Probability of picking a hiddenclass *h*,
P (*w* |*h*) = probability of generating a word.

Now we can formulate an observed pair (d, w)while the class variable h is eliminated. Theexpression computed after converting the wholeprocess into a join probabilistic model isexpressed as follows:

$$P(d, w) = P(d) * P(w | d), \ldots (1)$$

Where

$$P(w \mid d) = \sum P(t | h) * P(h | d) \ldots (2)$$

PLSA is an extension of LSA, so like LSA model andvector space model, input of the PLSA model is theword – document matrix X. This matrix X containingwords w ranges from 1 to m and documents d rangesfrom 1 to n and the total number of topic is H, to besought. X (w, d) represents the corresponding wordand document entry in specified row and column.Remembering the Random Sequence Model,referencing this model can show that:

$$P(d)= P(w_1 |d) * (w_2 |d)\ldots\ldots P(w_m |d)$$

$$mX (w, d) = P \prod (w_m |d), w = 1 \ldots (3)$$

If we have *H* topics as well:

$$P (w_m |d) = \Sigma P (w_m | topic_h) * P (topic_h | d), h = 1 \ldots (4)$$

The same written using shorthand:

$$P (w | d) = \Sigma P (w |h) *P (h |d), h = 1 \ldots (5)$$

So by replacing this, for any document in the collection, mX(w, d).

$$P (d) = \prod \{\Sigma P (w |h) *P (h |d)\}, w = 1\ h = 1 \ldots (6)$$

Now we found the two parameters for this model arep (w | h) and p (h | d).Here it is conceivable to infer the conditions for registering these parameters by Maximum Likelihood. After doing so we will get P (w |h) for allw and h, is a word by topic matrix (This gives thewords which make up topic).P (h |d) for all h and d, is a topic by document matrix(gives This gives the topic of document).The log likely hood of this model is the logprobability of the entire collection:

$$\Sigma \log P (d) = \Sigma X (w, d) \log \Sigma P (w | h) *P (h |d)$$

Where $d = 1$ $w = 1$ $h = 1$ ... (7)Which is to be maximized w.r.t. parameters P (w |h)and also P (h |d), subject to constraints thatS P (w |h) = 1 and S P (h |d) = 1 where $w = 1$ $h = 1$s.

EM algorithm consist two steps as follows:

1. In Expectation Step, current estimates ofparameters are used to compute posterior probability for hidden variables.

2. In Maximization-step, posterior probabilities thatare computed in Expectation steps are used to updateParameters.

The EM algorithm [10] is guaranteed to increase thelikelihood at each iteration. Following is the PLSAcalculation that accurately delineates legitimate info, preparing steps and yield given by this calculation.

Algorithm

Inputs: Word to document matrix T (w, d), $w = 1 : m$, $d = 1 : n$ and the number of topics sought.

Initialize arrays $P1$ and $P2$ randomly with numbers and normalize them row-wise.

Iterate until convergence.

For $d = 1$ to n, For $w = 1$ to m, For $h = 1$:

$P1 (w, h) = P1 (w, h) \Sigma \{X (w, d) * P2 (h, d) / \{ \Sigma P1 (w, h) * P2 (h, d)\}\}$

Where $d = 1$ $h = 1$ ... (8)

$P2 (h, d) = P2 (h, d) \Sigma \{X (w, d) * P1 (w, h) / \{\Sigma P1 (w, h) * P2 (h, d)\}\}$, $w = 1$ $h = 1$ ... (9)

$P1 (w, h) = P1 (w, h) / \Sigma P1 (w, h)$, $w = 1$ ... (10)

$P2 (h, d) = P2 (h, d) / \Sigma P2 (h, d)$ where $h = 1$ ... (11)

Output: Arrays $P1$ and $P2$, which hold the estimatedparameters $P (w |h)$ and $P (h |d)$ respectively [11].

RESULT ANALYSIS

Observation of PLSA performance shows that, whenone performs various tests to check the performanceof PLSA model, he/ she will defiantly get results thatare quite useful and appreciable also. PLSAcategorizes all next keywords according to sometopic and gives an extra edge to the query expansionfor specific domain.Some examples of PLSA results are illustrated infollowing tables:

Table 2 : Results of PLSA for query "Australian University"

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
University	Forum	Museum	Museum	AIU
Australian	Study	Forum	Forum	Buy
Australia	UNDA	England	Large	Security
ANU	JCU	Images	Books	Below
Research	CQU	Large	Architecture	Counter
Page	SCU	Above	Here	Sells
Student	CDU	Books	Churches	Login
International	ECU	Here	Images	Whistle blowing

Table 2 shows next keywords for the query“Australian University” and in results topic 1 simplyshows general term as “student”, “international”,

“ANU”, ”research” that are related to AustralianUniversity. Topic 2 contains terms like “UNDA”,“JCU”, “CQU”, “SCU”, “CDU”, “ECU” which areacronyms of respectively “University of Notre DameAustralia”, ”James Cook University”, “CentralQueensland University” and so on. Hence, secondtopic shows “List of Australian Universities”. In thesame way other topics can be easily understood.These terms can be used for query-expansion andwill in turn yield focused search.

OPTIMAL VALUES FOR NUMBER OF TOPICS (H)

The number of topics, ‘h’, in PLSA is one of the mostimportant factors. Its value must be an optimal one. Alarge value of ‘h’ will give some redundant topicsthat will not be informative enough and similarly asmall value will hide some useful concept. Results ofvarious tests suggest that this value should be inbetween 3 to7 for most of the cases of current Metasearchengines because at maximum level it will have24 to 27 documents. For such a specified number, therange of 3 to 7 topics is appropriate. An example forincreasing value of h is shown for same query “India
Tourism”. Every one of the terms in various themes are indicating distinctive viewpoints and criticalness.

Table 3: Results of PLSA for query “India Tourism” for different value of num of topic ‘a’=1, 2, 3

Topic1	Topic2	Topic3
India	Yimg	Kalpa
Tour	Directly	Demanding
Travels	JS	Manmade
Tourism	Hyatt	Munsiyari
Rajasthan	Marriot	Interzigm
Kerala	Regency	Wing

Convergence Behavior

Since PLSA uses EM for maximum likelihood, it alsoguarantees a convergent behavior for the iterativeprocedure. It always tries to find local maxima forgiven data distribution. PLSA also shows convergingbehavior in context for Meta search engine and wecan check it by using two measures named asfollows:

- Absolute Measure
- Average Measure

Absolute Measure

It can be computed by following formula

Maxi,j = | Pi,j
n+1 – Pi,j
n |

Where

Pi,j
n = value at ith
row and jth
column of word-topicmatrix or topic-document matrix after nthiteration.

In PLSA, firstly some random values are assigned toboth word-topic and topic-document matrix. Aftergoing through one iteration of the E and M steps, thealgorithm generates two new versions of thesematrices. This new version now acts as an input forthe next iteration of the algorithm and this iterativeprocedure continues till convergence. For measuringconvergence we compute the maximum differenceMaxi,j between all the corresponding cell entries ofword – document matrix and its newer version. Thiscalculation is performed for each iteration and themaximum value is noted

Average Measure

The average measure can be computed by the
following formula

Maxi,j = | Pi,j
n+1 – Pi,j
n | / 2 (|Pi,j
n+1 + Pi,j
n |)

Where

Pi,j
n = value at ith row and jth column of word-topic
matrix or topic-document matrix after nth iteration.The same procedure as previously explained, is usedhere. Only average measure is used in place ofabsolute measure.

APPLICATION OF PLSA

Performance of PLSA is observed better that that ofLSA model as the results of PLSA provide morerefined search results for given query. This is becausePLSA has solid statistical foundation. PLSA depends on the restrictive likelihood central and make utilization of EM calculation, or, in other words combine and thus deliver better outcomes. LSA has solutionfor the problem of synonymy only, but still after thesolution of synonymy polysemy is next problem thatis to be solved. PLSA solves both the problems veryefficiently. PLSA classify all the word to topicdistribution data in such a manner so that polysemousword is clubbed with other words with differentprobability and therefore represents different topics.In the previously explained example for query Indiatourism Aspect1 seems related to the places to visit inIndia as part of India tourism. Aspect2 tells aboutfamous hotels in India to stay for tourists. In othergroups all the famous hotel-name and restaurants as-“Hyatt”, “Marriot”, “Regency” are present whichrepresent another important aspect of “Tourism inIndia”. Aspect3 shows relevance with the restaurantsin India where visitors may go. PLSA is already inuse in some applications and contributing fruitfulresults. Apart from already explained domain whererelevant document are retrieved for given query;PLSA is used in “Web Page Clustering using PLSA”and in the “Multimodal Image Retrieval usingPLSA “.

CONCLUSION AND FUTURE WORK

In this paper we have reviewed how meta searchengine produces results, on which principal they arebased and also we study that the result produced byMeta search engine are refined up to a desired levelor not. After doing various experiments with searchapproach we come to the point and concluded thatPLSA can provide efficient result for queryexpansion. In these experiments we saw that PLSAperforms better that previously used technique i.e.LSA and produces all the results in well-classifiedand easily understandable form. In future we canmodify our approach with the use of a new systemthat is called “Named Entity Recognizer” in MSE.

REFERENCES

[1] Effective Internet Search Strategies: InternetSearch Engines, Meta-Indexes, and WebDirectories, by Wendy E. Moore, M.S. in L.S.,Acquisitions/Serials Librarian, The University of
Georgia School of Law Library Athens, GA

[2] Shanmukha Rao B.; Rao S.V.; Sajith G.; “AUser-profile Assisted Meta Search Engine”,
TENCON 2003 Conference on ConvergentTechnologies for Asia-Pacific Region, 2, pp. 713
– 717, 15 – 17 Oct. 2003.

[3] Spink A.; Jansen B.J.; Blakely C.; Koshman S.;“Overlap Among Major Web Search Engines”,
ITNG 2006 Third International Conference onInformation Technology.

[4]A Vector Space Model for automatic indexinggiven by G. Salton, Cornell Univ. Ithaca, NYA.
WongCornell Univ., A. WongCornell Univ., Ithaca, NYC. S. YangCornell Univ., Ithaca,
NY, Magazine: Communications of the ACM CACM Homepage archive, Volume 18
Issue 11, Nov. 1975 , Pages 613 – 620.

[5] Query Optimization Using Genetic Algorithms inthe Vector Space Model by Eman Al Mashagba,Feras Al Mashagba, Mohammad Othman Nassarin IJCSI International Journal of ComputerScience Issues, Vol. 8, Issue 5, No 3, September2011 ISSN (Online): 1694-0814 www.IJCSI.org.
[6] Hoffmann, T.: Collaborative Filtering viaGaussian Probabilistic Latent Semantic Analysis. In: Proceedings of the 26th annual internationalACM SIGIR conference on Research and development in information retrieval, ACMPress (2003) 259-266.
[7] Hofmann, T.: Probabilistic Latent SemanticAnalysis. In: Proceedings of Uncertainty in Artificial Intelligence, UAI'99, Stockholm(1999).[8] Dempster, A.P., Laird, N.M., Rubin, D.B.: (Maximum likelihood from incomplete data viathe EM algorithm.
[9] Prof. Pangfeng Liu gives PLSA Search Engine,2008 Parallel Programming, Department of Computer Science and Information Engineering,National Taiwan University.
[10] Jie Xu, Getian Ye, Yang Wang, GunawanHerman, Bang Zhang, Jun YangNational ICT Australia School of ComputerScience and Engineering, University of NewSouth Wales, Incremental EM for ProbabilisticLatent Semantic Analysis on Human ActionRecognition.
[11] International Journal of Electronics Engineering,2 (2), 2010, pp. 381 – 384 A Framework for
Analysis of the Applicability of ProbabilisticLatent Semantic Analysis Technique in MetaSearch Engine.
[12] VijayshareeGautamImproving Search Strategy of Search Engine Using Probabilistic Latent SemanticAnalysis Technique