IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Improving Search Strategy of Search Engine Using Probabilistic Latent Semantic Analysis Technique

Vijaysharee Gautam[*1], Devendra Kumar Sharma[2], Vaishali Shringi[3]
[*1]M. Tech. Scholar, Suresh Gyan Vihar University, Jaipur, Rajasthan.
[2,3]M. Tech. Scholar, Rajasthan Technical University, Kota, Rajasthan
vijayshreegautam@gmail.com

Abstract

Users on the internet uses search engine to find information of their interest. However current search engines on web return answer to a query of user independent of user's requirement for the information. In this paper our aim is to use a new technique called probabilistic latent semantic analysis which gives search results to user that are more accurate than previously used techniques by various search engines. Our main focus in this paper is on the requirement for more accurate search results by meta search engine. In comparison with previously used techniques, for searching like LSA, which perform singular value decomposition of co-occurrence tables, proposed technique of this paper, relies on mixture decomposition derived from latent class model. Results obtained by PLSA in searching for query shows that this technique gives more accurate results in searching most relevant document from a given corpus for a query of user.

Keyword - Meta search engine, Indexing Query, Vector Space Model, Latent Semantic Indexing, Word-Document Matrices, Probabilistic Latent Semantic Analysis, Expectation Maximization algorithm.

Introduction

In recent years internet is growing rapidly and its popularity is increased up to a point that every person knows about it and make its use for different - different purposes. Some people use internet to know about something new in the current environment while others use it as a means of entertainment. The use of internet is not limited to entertainment but it can also be used to conduct research related work, like finding and reading latest researches on current trends. Internet is also used for getting latest news. Millions of web pages are added to this internet with all human needs. A survey by Google represent that there are one trillion unique URL's on the internet. The implementation of search engine makes the process of searching some of the topics of user interest in a easy way. Queering the search engine for any particular topic would retrieve the results from the internet and presented to the web users. Since there are large number of web pages on the internet and thus result obtained are also vast. User gets more than enough web links as a result produced by search engine and wastes their precious time in navigating through unwanted links, searching the needed one. The main reason for this is that the Search Engine do the indexing of the pages on the basis of text entered by user. In order to overcome this shortcoming we need to implement a method that will allow the user to find the relevant words, starting from the few words that they may actually know. In other words, we need to focus on the semantic of words entered by user. This research paper presents a new approach that is based on some algorithms which considers semantic aspects of text and uses them to implement a Meta search engine that will give user appropriate results in their query for relevant information. For computers to interact more naturally with humans, It is necessary to deal with users requests that do not have clear meaning or we can say that deal with impractical user requests is necessary. It is a important need to recognize the difference between what a user might say or do and what he/she actually want and intended for.

A process of information retrieval using web search engines comprises following steps.

1. By natural language processing, for e.g. user provides some keywords to web search engine and expects that it will return the 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 special program saves all the words and their respective position on the visited web-page.
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 different search

engine for indexing is different and thus the result produced by different search engine for the same query is different. Important points considered during indexing process include: the frequency of a term appearing in a web-page, portion of a web-page where that term appears, font-size of a term.

Indexing information is encoded into reduced size to speed up the response time of particular search engine, and then it is stored into the database.

Problem Definition And Statement

Techniques adopted for by meta search engine in searching a document relevant to user query not giving the satisfactory results to the users. The principal behind the techniques used is either extracting the preferences given by user or maintaining user profile. Some iterative algorithms are applied on search engine results for the purpose of refining the results appropriately and more accurately. Output of these algorithms gives the solution of the problem definition explained here in this section. The main point of thinking is the choice of appropriate algorithm for improving the search strategy of Meta search engine [1]. When working with search engine users faces a common problem of not getting the desired information quickly in an easy way. The main problem is that when user enters some text keyword in search engine, it will return a list of various web pages on the basis of keyword typed by the user. Usually search engine does not respond with only the result that user actually needed, instead it gives lots of undesirable web page links and user wastes their precious time in navigating from one web page to another in search for the document what they actually want.

For improving the search strategy keywords typed by the user in search for the information what they needed is also an important issue. Many internet users want information of their interest on web, but they do not know how to get that information fast in a an easy way. The choice of keyword typed by user is also a critical issue. Another aspect of problem definition depends on the ability of search engine to respond with appropriate search result. Not any search engine discovered yet, is capable of covering even a half portion of the web pages available on the net [2]. Some search engines gives the web pages that are visited many times and thus the required page does not come in front of the user and they make search again and again, but always gets the same result for a given keyword through a specific search engine. An even sometimes search engine gives such web page links in results which contain obsolete or dead link [3].

A study was performed to evaluate the similarities and differences between the search results given by the three search engines named Google, Yahoo, AskJeeves, and this process is named evaluating overlapping among first page results of the above mentioned search engines. This analysis reveals that 92.53 percent of URL are retrieved by one search engine only (which could be any out of the three) , 5.22 percent URLs are shared by two, while 2.02 percent and 0.21 percent of URLs were retrieved by all three search engines. This small percentage of overlapping between SEs shows that there is a significant difference in search strategy of all SEs. Page ranking and retrieval methods adopted by all SEs are found different as a result of overlapping survey. Now with this survey one can say that if internet user uses only one search engine in search for the document of theirs interest , then they will not get the desired result because it is not sure that the SE, they are using will defiantly give them the required document and they may miss needed and relevant data . Interacting with only one search engine and not getting help with the remaining one can make user loss in search for information, Because the search strategy used by all SEs are different.

Proposed Framework

Our proposed model is based on the Vector Space Model and later we further extend it to the PLSA (Probabilistic Latent Semantic Analysis) model and then examine how these models worked to perform query expansion. On Internet various text retrieval techniques are based on indexing of text keyword, since keyword alone is not capable of capturing the whole document content appropriately, the performance of retrieval strategy becomes poor. Bu t using the indexing mechanism of keywords we can process large corpa of document in an efficient way. When identification of significant index word is finished one of the two information retrieval model is used to match query to document named statistical model or Boolean model. Statistical model gives the similarities between query and document while Boolean model matches to an extent up to which the word satisfies Boolean expression. In 1975 Gerald Slaton [4] gives a model named “Vector Space Model” which maps the document in n-dimensional space. Where n is the number of different words ($w_{1,}$ $w_{2,}$ w_3 ....w_n) which contains the whole vocabulary of the the corpus or text collection. Each dimension corresponds to a separate term. If a term exits in the document, its value in the vector is non zero. Vector operations can be used to compare document with queries. In vector space model each document is considered as a vector as $D_{1,}$ $D_{2,}$ $D_{3,}$ $D_{4,}$,............ $D_{r,}$ Where r is the total number of document in corpa.

Representation of document vector is

$$Dir = (d_{1r,}\ d_{2r,}\ d_{3r,}.......................\ ,\ d_{nr})$$

dir represents the i^{th} component of r^{th} document vector.

Concept of Vector Space Model

Vector Space Model is an algebraic model for representing text documents as vectors of identifiers. It is used in information filtering. Traditionally this model is used where documents are placed in term – space. Query is also like a very short document. This model is required to find the most relevant document for the given query. In this model computation of similarities between collection of documents and query is performed first and then returns the most accurately matching documents [4]. Similarities are computed on basis of various different factors. One of them, very frequently used similarity factor is the cosine similarity. Similarity between document vector and query vector can be calculated by, comparing the deviation of angles between each document vector and the original query vector.

In practice it is easier to calculate the cosine of angle between the vectors, instead of angles itself.

$\cos \square = Q*D/|Q|*|D|$

The expression shows the cosine angles between document vector D and query vector Q. If two documents are neighbors of each other in term space then they would be considered relevant with each other. By applying different similarity measures to compare queries to terms and documents, properties of the document collection can be emphasized or deemphasized. For example dot product similarity measure finds the Euclidean distance between the query and a term or document in the space. Also the cosine similarity is mentioned above. Here some other factors are also mentioned for measuring similarity between document vector and query vector [5].

Table 1: Similarity measures of VSM

Similarity Measure	**Evaluation of binary term vector**
Cosine similarity	$\cos\theta = Q*D/\lvert Q\rvert*\lvert D\rvert$
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 associated with numeric factor and that numeric factor is called weight of the respective word or term in document. Weight associated with word $w_{i,}$ can be replaced by term frequency (tf_i).

Here some advantages of Vector Space model over Boolean model are listed below.

1. VMS is a simple model based on linear algebra.
2. Term weights not binary.
3. VMS allows for calculating a continuous degree of similarity between queries and documents.
4. It allows ranking of documents based on their possible relevance.

Some limitations of VMS are mentioned below.

1. Long document are poorly represented because they have poor similarity values.
2. Search keywords must precisely match document terms.
3. Semantic sensitivity: documents with similar context but different term vocabulary won't be associated, resulting in a false negative match.
4. The order of term appearing in the document has lost in in vector space representation.
5. Weighting is intuitive but not very formal.

Concept of Proposed Technique(PLSA)

Th. Hofmann presented a statistical view on LSA, which formulate the new model called Probabilistic Latent Semantics Analysis model [6][7], which provide probabilistic approach for discovering latent variables, which has a statistical foundation. The basic of PLSA is a latent class statistical mixture model named Aspect model. This aspect model assumes that there is a set of hidden factors underlying the co occurrences between two documents. PLSA uses Expectation-Maximization (EM) [8] to estimate the probability values that measure the relationship between the hidden factors and the two sets of documents. In this model we represents the hidden class variable h € H = $\{h_{1,}\ h_{2,}\ h_{3,}...........\}$, document d € D = $\{\ d_{1,}\ d_{2,}\ d_{3,}...........\}$ and words w € W = $\{w_{1,}\ w_{2,}\ w_{3,}...........\}$.

Some parameters of this model can be defined in the following way [9]:

P (*d*) = Probability of selecting a document *d*,

P (h |*d*) = Probability of picking a hidden class *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. The expression computed after converting the whole process into a join probabilistic model is expressed as follows:

$P\ (d, w) = P\ (d) * P\ (w \mid d)$, ... (1)

Where

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

PLSA is an extension of LSA, so like LSA model and vector space model, input of the PLSA model is the word – document matrix X. This matrix X containing words w ranges from 1 to m and documents d ranges from 1 to n and the total number of topic is H, to be sought. X(w, d) represents the corresponding word and document entry in specified row and column.

Remembering the Random Sequence Model, Referencing this model can show that:

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

$mX\ (w, d) = P \prod (w_m \mid d), w = 1$ ... (3)

If we have H topics as well:

$P\ (w_m \mid d) = \Sigma P\ (w_m \mid \text{topic}_h) * P\ (\text{topic}_h \mid d), h = 1$ ... (4)

The same written using shorthand:

$P\ (w \mid d) = \Sigma P\ (w \mid h) * P\ (h \mid d), h = 1$ ... (5)

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

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

Now we found the two parameters for this model are p (w | h) and p (h | d).

Here it is possible to derive the equations for computing these parameters by Maximum Likelihood. After doing so we will get P (w |h) for all w and h, is a word by topic matrix (This gives the words 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 log probability of the entire collection:

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

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 that

$\Sigma P\ (w \mid h) = 1$ and $\Sigma P\ (h \mid d) = 1$ where $w = 1\ h = 1$

s

EM algorithm consist two steps as follows:

1. In Expectation Step, current estimates of parameters are used to compute posterior probability for hidden variables.
2. In Maximization-step, posterior probabilities that are computed in Expectation steps are used to update Parameters.

The EM algorithm [10] is guaranteed to increase the likelihood at each iteration. Following is the PLSA algorithm that precisely depicts proper input, processing steps and output given by this algorithm.

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 estimated parameters P (w |h) and P (h |d) respectively [11].

Result Analysis

Observation of PLSA performance shows that, when one performs various tests to check the performance of PLSA model, he/ she will defiantly get results that are quite useful and appreciable also. PLSA categorizes all next keywords according to some topic and gives an extra edge to the query expansion for specific domain.

Some examples of PLSA results are illustrated in following 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 simply shows general term as "student", "international", “ANU”, ”research” that are related to Australian University. Topic 2 contains terms like “UNDA”, “JCU”, “CQU”, “SCU”, “CDU”, “ECU” which are acronyms of respectively “University of Notre Dame Australia”, ”James Cook University”, “Central Queensland University” and so on. Hence, second topic shows “List of Australian Universities”. In the same way other topics can be easily understood. These terms can be used for query-expansion and will in turn yield focused search.

Optimal Values For Number Of Topics (h)

The number of topics, ‘h’, in PLSA is one of the most important factors. Its value must be an optimal one. A large value of ‘h’ will give some redundant topics that will not be informative enough and similarly a small value will hide some useful concept. Results of various tests suggest that this value should be in between 3 to7 for most of the cases of current Meta-search engines because at maximum level it will have 24 to 27 documents. For such a specified number, the range of 3 to 7 topics is appropriate. An example for increasing value of h is shown for same query “India Tourism”. All the terms in different topics are showing different aspects and significance.

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 also guarantees a convergent behavior for the iterative procedure. It always tries to find local maxima for given data distribution. PLSA also shows converging behavior in context for Meta search engine and we can check it by using two measures named as follows:

- Absolute Measure
- Average Measure

Absolute Measure

It can be computed by following formula

$$Max_{i,j} = | P_{i,j}^{n+1} – P_{i,j}^{n} |$$

Where

$P_{i,j}^{n}$ = value at ith row and jth column of word-topic matrix or topic-document matrix after nth iteration.

In PLSA, firstly some random values are assigned to both word-topic and topic-document matrix. After going through one iteration of the E and M steps, the algorithm generates two new versions of these matrices. This new version now acts as an input for the next iteration of the algorithm and this iterative procedure continues till convergence. For measuring convergence we compute the maximum difference $Max_{i,j}$ between all the corresponding cell entries of word – document matrix and its newer version. This calculation is performed for each iteration and the maximum value is noted

Average Measure

The average measure can be computed by the following formula

$$Max_{i,j} = | P_{i,j}^{n+1} – P_{i,j}^{n} | / 2 (| P_{i,j}^{n+1} + P_{i,j}^{n} |)$$

Where

$P_{i,j}^{n}$ = 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 used here. Only average measure is used in place of absolute measure.

Application of PLSA

Performance of PLSA is observed better that that of LSA model as the results of PLSA provide more refined search results for given query. This is because PLSA has solid statistical foundation. PLSA is based on the conditional probability principal and make use of EM algorithm, which is guaranteed to converge and hence produce better results. LSA has solution for the problem of synonymy only, but still after the solution of synonymy polysemy is next problem that is to be solved. PLSA solves both the problems very efficiently. PLSA classify all the word to topic distribution data in such a manner so that polysemous word is clubbed with other words with different probability and therefore represents different topics. In the previously explained example for query India tourism Aspect1 seems related to the places to visit in India as part of India tourism. Aspect2 tells about famous hotels in India to stay for tourists. In other groups all the famous hotel-name and restaurants as- “Hyatt”, “Marriot”, “Regency” are present which represent another important aspect of “Tourism in India”. Aspect3 shows relevance with the restaurants in India where visitors may go. PLSA is already in use in some applications and contributing fruitful results. Apart from already explained domain where relevant document are retrieved for given query; PLSA is used in “Web Page Clustering using PLSA” and in the "Multimodal Image Retrieval using PLSA".

Conclusion And Future Work

In this paper we have reviewed how meta search engine produces results, on which principal they are based and also we study that the result produced by Meta search engine are refined up to a desired level or not. After doing various experiments with search approach we come to the point and concluded that PLSA can provide efficient result for query expansion. In these experiments we saw that PLSA performs better that previously used technique i.e. LSA and produces all the results in well-classified and easily understandable form. In future we can modify our approach with the use of a new system that is called "Named Entity Recognizer" in MSE.

References

[1] Effective Internet Search Strategies: Internet Search Engines, Meta-Indexes, and Web Directories, 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.; "A User-profile Assisted Meta Search Engine", TENCON 2003 Conference on Convergent Technologies 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 on Information Technology.

[4]A Vector Space Model for automatic indexing given 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 in the Vector Space Model by Eman Al Mashagba, Feras Al Mashagba, Mohammad Othman Nassar in IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3, September 2011 ISSN (Online): 1694-0814 www.IJCSI.org.

[6] Hoffmann, T.: Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press (2003) 259-266.

[7] Hofmann, T.: Probabilistic Latent Semantic Analysis. 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 via the 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, Gunawan Herman, Bang Zhang, Jun Yang National ICT Australia School of Computer Science and Engineering, University of New South Wales, Incremental EM for Probabilistic Latent Semantic Analysis on Human Action Recognition.

[11] International Journal of Electronics Engineering, 2 (2), 2010, pp. 381 – 384 A Framework for Analysis of the Applicability of Probabilistic Latent Semantic Analysis Technique in Meta Search Engine.