



STUDY OF INDEXING TECHNIQUES FOR WEB CRAWLED DATA

Vaishnave. M. R, Sreepriya. V, Aravind. K, Bindu.K.R

Department of Information Technology, Amrita Vishwa Vidyapeetham, Coimbatore, India.

Abstract: Data collected from web are unclean, amorphous, formless and unstructured. In order to get the information from the collected data, it has to be cleaned, structured and indexed. There are different indexing techniques, each having its own advantages and disadvantages say for instance LSI is less complex but has scalability problems. Depending on the purpose appropriate indexing technique has to be chosen. This paper deals with the study of Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Indexing (PLSI) and Regularized Latent Semantic Indexing (RLSI).

Keywords: Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Indexing (PLSI) and Regularized Latent Semantic Indexing (RLSI).

I. INTRODUCTION

The amount of data in today's web is vast and finding the necessary resources can be tedious even for a machine, let alone man. The data necessary should not only be identified but should also be found in the necessary format and needed time as data may lose its value as it ages. This is where search engines come into play. They use in-built crawlers to crawl the web for information. But the data tend to be noisy and difficult which in turn leads to significant wastage in amount of network and local resources. The data retrieved is to be cleaned, integrated and transformed to tailor the needs. This data is then indexed and stored in databases and retrieved based on input query. Earlier indexing concepts being LSI suffered certain shortcomings when size of the data increased. New indexing techniques came on to overcome the shortcoming of the previous ones.

II. LATENT SEMANTIC INDEXING

Before LSI, classic Information Retrieval (IR) was used. But classic IR had lower performance because its answer set might include all related documents where all may not be relevant to a particular topic at that particular point of time. In order to retrieve a relevant document in classic IR it should contain at least one index term, which is not possible always and retrieval based on index will be vague and noisy. Generally the user's information need is more related to concepts and ideas than to index terms and a document that shares concepts with another document known to be relevant which might be of interest. The key idea is to map documents and queries into a lower dimensional space that is composed of higher level concepts which are in fewer number than the index terms. Documents are represented as "bags of words", where the order of the words in a document is not important, only how many times each word appears in a document. Concepts are represented as patterns of words that usually appear together in documents and words are assumed to have only one meaning. Latent Semantic Indexing (LSI) uses Singular Value Decomposition (SVD) and creates a term-document matrix to retrieve documents based on search query. Retrieval in this reduced concept space might be superior to retrieval in the space of index terms. LSI reduces the complexity problem imposed by classic IR by considering semantically related terms rather than all the related terms. But LSI can generate negative values and the size of the matrix can increase in the order $O(\min\{MN^2, NM^2\})$. This is why the next technique of PLSI was developed.

III. PROBABILISTIC LATENT SEMANTIC INDEXING

PLSI has a more robust statistical foundation and provides a proper generative data model and it uses the EM algorithm. PLSI was found more effective when compared to LSI because in LSI it analyses the co-occurrences of words and documents and it forms a term by topic matrix i.e. it gives what terms gives a topic and which topics are in which documents. It uses topics to create queries and uses the Gaussian model which can generate negative values but negative number of words in a document is not possible. It uses a multi-nominal model i.e. probability distribution. Documents can be represented as numeric vectors in the space of words. The order of words may be lost but the co-occurrences of words still provides insight about the topic in a collection of document. It is a good basis for information retrieval system since it helps to find what topics are there in a collection of documents and helps to overcome the problems in LSI. These help us to have a more robust statistical foundation and provides generative data model. Since it can deal with domain specific synonymy and polysemous words, it is more flexible. It is an unsupervised method. PLSI technique maintains probability

distribution constraints of term-topic matrix. So to overcome these disadvantages a recent method of indexing is developed, Regularized Latent Semantic Indexing (RLSI).

IV. REGULARISED LATENT SEMANTIC INDEXING

It differs from LSI in the way that it uses Regularisation rather than Orthogonality. Topic model represents the relationship between terms and document through latent topics. This can be formalised as minimisation of quadratic loss function on term-document occurrences regularised by l1 or l2 norm. RLSI decomposes the problem into multiple sub-optimization problems which may run in parallel thereby overcomes the scalability problem posed by the previous techniques. The indexed data is stored into a database from which the data is retrieved based on the query specified and presented to the user using desired application in the front end. In the term-document matrix, each entry is occurrence of term in document. The term-document matrix is approximation of product of term-topic (latent topics with terms) and topic-document matrix (document with topics). The formulation of RLSI makes it possible to decompose the problem into multiple sub-optimisation problems which may run in parallel. This is the main reason why RLSI can scale up. When the number of terms is large the problem of scalability becomes particularly severe. Regularization is a common technique used in machine learning to prevent this over-fitting. RLSI can also be applied for relevance ranking in Information Retrieval (IR) field to reduce the term mismatch. Term mismatch arises when authors of document and users of search use different terms to denote the same concept. In the topic space it is likely that the two terms are in the same topic and using of matching score in topic space along with term matching scores improves relevance ranking.

V. CONCLUSION

From the various methods discussed above, we can conclude that LSI can be used in cases wherein improve the efficiency of the retrieval process, decreases the dimensionality of vectors. Good for machine learning algorithms in which high dimensionality is a problem and dimensions are more semantic. RLSI is preferred over other indexing techniques such as LSI, LDA in terms of scalability, readability, coherence and performance.

REFERENCES

- [1] Hofmann, T. Probabilistic latent semantic analysis. In Proceedings of the 15th Conference on Uncertainty in AI (1999).
- [2] Regularized Latent Semantic Indexing by Quan Wang, Jun Xu ,Hang Li and Nick
- [3] Indexing by Latent Semantic Analysis Scott Deerwester; Susan T Dumais; George W Furnas; Thomas K Landauer Richard. Journal of the American Society for Information Science (1986-1998);